Multigene datasets for deep phylogeny

Martin Jones

Thesis presented in accordance with the requirements for the

degree of Doctor of Philosophy

University of Edinburgh

2006

Declaration

I declare that this thesis has been composed by myself and, except where otherwise stated, is entirely my own work.

Martin Jones

December 2006

Contributions of co-authors

The work described here has formed the basis of two papers:

Jones, M; Gantenbein, B; Fet, V and Blaxter, M. The effect of model choice on phylogenetic inference using mitochondrial sequence data: lessons from the scorpions. In press Molecular Phylogenetics and Evolution, November 2006.

Martin Jones carried out the data gathering and analysis and wrote the paper. Mark Blaxter assisted with interpretation of the results and supervised the project. Benjamin Gantenbein and Victor Fett sequenced the *Mesobuthus gibbosus* mitochondrial genome. All authors assisted with the drafting of the paper.

Jones, M and Blaxter, M. TaxMan: a Taxonomic database Manager. Accepted BMC Bioinformatics, December 2006.

Martin Jones wrote the software and manuscript. Mark Blaxter assisted with software design and testing and supervised the project.

Abstract

Though molecular phylogenetics has been very successful in reconstructing the evolutionary history of species, some phylogenies, particularly those involving ancient events, have proven difficult to resolve. One approach to improving the resolution of deep phylogenies is to increase the amount of data by including multiple genes assembled from public sequence databases. Using modern phylogenetic methods and abundant computing power, the vast amount of sequence data available in public databases can be brought to bear on difficult phylogenetic problems.

In this thesis I outline the motivation for assembling large multigene datasets and lay out the obstacles associated with doing so. I discuss the various methods by which these obstacles can be overcome and describe a bioinformatics solution, TaxMan, that can be used to rapidly assemble very large datasets of aligned genes in a largely automated fashion. I also explain the design and features of TaxMan from a biological standpoint and present the results of benchmarking studies. I illustrate the use of TaxMan to assemble large multigene datasets for two groups of taxa – the subphylum Chelicerata and the superphylum Lophotrochozoa.

Chelicerata is a diverse group of arthropods with an uncertain phylogeny. When a set of mitochondrial genes is used to analyse the relationships between the chelicerate orders, the conclusions are highly dependent upon the evolutionary model used and are affected by the presence of systematic compsitional bias in mitochondrial genomes.

Lophotrochozoa is a recently-proposed group of protostome phyla. A number of distinct phylogenetic hypotheses concerning the relationships between lophotrochozoan phyla have been proposed. I compare the phylogenetic conclusions given by analysis of nuclear and mitochondrial protein-coding and rRNA genes to evaluate support for some of these hypotheses.

The multigene approach to phylogenetics holds great promise for addressing previously intractable problems. While the availability of both computing power and sequence data looks bound to increase, the applicability of the multigene strategy will be limited by the sophistication of bioinformatics tools.

Acknowledgements

There are many people I'd like to thank for a huge variety of reasons.

Firstly, my parents have given me unceasing support in everything I've done. None of my achievements would have happened without them.

My colleagues at Edinburgh, both past and present, have been wonderful and a pleasure to work with. In no particular order: Ann, Alasdair, Jenna, Seanna, Ralf, Claire, Habib, Fran, and Katelyn all deserve my thanks. Special thanks go to James for proof-reading duties. Mark, my supervisor, have given me all his support and a great environment in which to work. My co-authors Victor and Benjamin have helped me publish the work in this thesis. Thanks also go to Chris Jiggins for assisting in software testing and to Graham Stone for co-supervision and support.

Over the course of my PhD I've had the pleasure of living with a procession of great flatmates and friends. I want to thank Graham and Lea for their friendship over the last seven years, and for putting up with my messy ways.

This work was supported by a BBSRC studentship. Thanks are also due to the many people working on open-source software, without which this work could not have been carried out.

The biggest of hugs goes to Jane for making the last six months of my PhD happier than I ever thought they would be. She has had endless understanding and patience whenever it has been required, and I'm very lucky to have her.

"A tree's a tree. How many more do you need to look at?"

- Ronald Reagan, 1965 (Attrib.)

Brief Table of Contents

Abbreviations used

AIC : Akaike Information Criterion

APC : Acoelomata / Pseudocoelomata / Coelomata

AT : Adenine / Thymine

BF : Bayes Factor

BLAST : Basic Local Alignment Search Tool

CDS : Coding DNA Sequence

CG : Cytosine / Guanine

CPU : Central Processing Unit

EST : Expressed Sequence Tag

FTP : File Transfer Protocol

G : Gamma rate variation

GHz ; Gigahertz

GTR : General Time Reversible

I : Invariant sites

JC : Jukes-Cantor

LBA : Long Branch Attraction

LED : Lophotrochozoa / Ecdysozoa / Deuterostomia

LSU : Large Subunit

MB : Megabytes

MCMCMC : Metropolis Coupled Markov Chain Monte Carlo

ML : Maximum Likelihood

MR : Majority Rule

NCBI : National Centre for Biotechnology Information

NTE : Neutral Transitions Excluded

RAM : Random Access Memory

POA : Partial Order Alignment

RDBMS : Relational Database Management System

R/Y : Purine / Pyrimidine

SSU : Small Subunit

Table of Contents

Figure Index

Index of Tables

1 Introduction

One of the major impacts of the widespread use of DNA sequencing technology has been the use of molecular sequence data for phylogenetics. Nucleotides and amino acids make ideal characters for phylogenetic reconstruction, avoiding many of the pitfalls associated with phylogenetic analysis of morphological characters. Since the field's inception there have been numerous methodological improvements, most notably in tree reconstruction routines but also in related areas such as evolutionary model testing, dating of speciation events and phylogenetic hypothesis testing. Coupled with these improvements has been a rapid increase in the availability of computing power, with the result that the most sophisticated types of analyses, although much more computationally intensive than older methods, can be executed in a reasonable amount of time.

In parallel with more powerful methods and computers has come an exponential increase in the amount of sequence data available for analysis in freely accessible databases ("public sequence data"; Figure1.1, Benson *et al.* 2006).

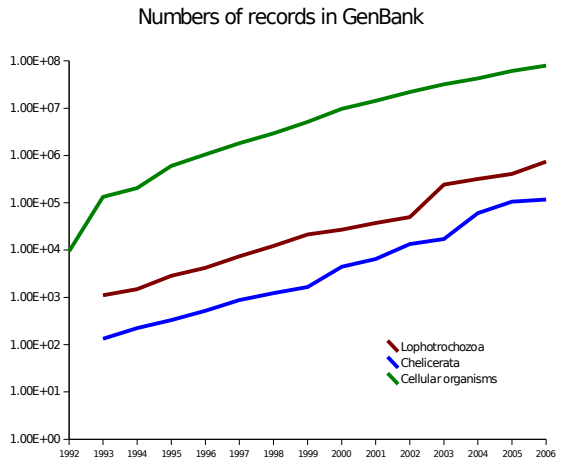


Figure 1.1: Growth of the GenBank sequence database

Lines show the parallel growth in numbers of records in GenBank for Lophotrochozoa, Chelicerata and all cellular organisms. Years are shown on the X-axis; the Y-axis shows number of records on a logarithmic scale.

Because the raw data required for molecular phylogenetic analysis consists simply of aligned orthologous nucleotides or amino acid residues, gene sequences that have been generated for purposes completely unrelated to phylogenetics can potentially be used, making public sequence databases a source of large quantities of phylogenetic information. The challenge for researchers is to develop methods of accessing the information contained in such databases which, as a rule, have not been designed with the phylogenetic inference in mind. As a result, a recent movement in phylogenetics has been the development of methods for large-scale phylogenetic analysis of multiple genes, in the hope of turning this bounty of sequence data into systematic knowledge (Hassanin 2006; Philippe, Lartillot and Brinkmann 2005; Rokas *et al.* 2003). Since

traditional manual approaches to sequence gathering are insufficient to cope with the massive amounts of data available, successful use of public sequence data for phylogenetics has required the use of bioinformatics approaches. It is only in a bioinformatics framework that the various components of multigene phylogenetics – modern phylogenetic methods, ample computing power, and large sequence databases – can be brought to bear on difficult phylogenetic problems. This project and thesis deals with species phylogenetics, the aim of which is to discern relationships between species and higher taxa. Additional fields of phylogenetic analysis (within-species analyses, investigation of gene families) are also facilitated by these advances.

1.1 Why do multigene phylogenetics?

Phylogenetic analysis of single genes has proven extremely effective in many cases, resolving relationships between species with much greater robustness than could be obtained with morphological characters. Relationships between very distantly related taxonomic groups (eg. phyla) become very difficult to resolve with morphological characters as character homology becomes more difficult to define between dissimilar organisms. However, several well-studied genes are sufficiently conserved across large taxonomic groups to make them amenable to phylogenetic analysis, and pioneering work has been carried out using genes that are present in all domains of life (Baldauf and Palmer 1993; Steenkamp, Wright and Baldauf 2006). However, attempts to use such genes for deep phylogeny has not been straightforward, and several scenarios have been outlined in which single genes may fail to resolve a phylogeny.

Evolutionary rate is a critical issue in phylogenetics. All molecular phylogenetic

methods attempt to use observed differences between sequences to derive relationships. A highly conserved gene with a very low rate of evolution may differ too little between closely-related taxa, or taxa which diverged over a relatively short period, to offer any phylogenetic signal on which tree reconstruction methods can work. Conversely, a highly variable gene with a rapid pattern of evolution may accumulate so many differences between taxa that phylogenetic signal is drowned out by noise and character homology may be very difficult to assign with confidence. This is particularly likely when the taxa involved are very distantly related. Even if phylogenetic signal is preserved, fast-evolving genes can exacerbate phylogenetic artefacts such as long branch attraction. To obtain wide-ranging phylogenies, relationships have to be resolved at multiple taxonomic levels. This is likely to be problematic when using a single gene – a gene offering good resolution of relationships between closely related species is unlikely to perform well for distantly related species, and yet this is exactly what is required for fully resolved trees of deep phylogenies.

Shorter genes, even if their rate of evolution is appropriate for the degree of relatedness of the taxa under investigation, may contain too few phylogenetically informative characters to allow a robust phylogeny to be determined. Methods of assessing support for a phylogenetic hypothesis include the bootstrap (Felsenstein 1985), the jackknife (Farris *et al.* 1996), and estimation of posterior probabilities (Erixon *et al.* 2003; Huelsenbeck and Rannala 2004; Svennblad *et al.* 2006). A gene may carry phylogenetic information that supports a given phylogeny, but if the phylogenetic signal is weak due to a small number of informative characters, confidence in the

conclusion, as determined by these methods, will be low. This is particularly likely to be the case for short branches which represent a small evolutionary distance.

When single genes are used for phylogenetic analysis it is on the assumption that they are representative of the evolutionary history of the genome as a whole. However, individual genes can have many different aberrant patterns of evolution which falsify this assumption and make them unsuitable for phylogenetic analysis. A gene can undergo accelerated evolution in a given lineage, or acquire a different pattern of compositional bias. Different positions in a gene can evolve differently between lineages. Any of these mean that the gene is not best suited for phylogenetic analysis of taxa, yet if analysis is carried out on such a gene, unambiguous (but incorrect) conclusions may be drawn. Analyses which use a single gene to infer species phylogeny are also vulnerable to being mislead by comparison of paralogous sequences, which are related by gene duplication rather than speciation.

1.2 Multiple gene phylogenies can avoid these problems

Using multiple genes for phylogenetics allows researchers to avoid the problems described above. In a study involving multiple genes, genes with different rates of evolution can be included in an analysis, increasing the prospects for resolution of relationships at all levels. Because of the greater number of characters available for analysis, phylogenetic conclusions are likely to be supported more robustly. This allows short branches to resolved with greater confidence. By summarising the phylogenetic information present in multiple genes, it is less likely that the conclusions

will be affected by abnormal patterns of evolution in a single gene. Additionally, one can test for the presence of different types of bias in a multigene dataset by systematically excluding genes with certain properties. For instance, Philippe, Lartillot and Brinkmann (2005) found that, in a large dataset of 146 genes, progressively removing the fastest-evolving genes led to an increase in bootstrap support for the Ecdysozoa hypothesis from ~0.2 to ~0.9, dramatically changing the conclusions of the study. This represented the signature of a long branch attraction effect: had one of the fast-evolving genes been used in a single gene study, erroneous conclusions would have been drawn.

Supertrees and Supermatrices

Two competing methods exist for combining information from multiple genes, commonly referred to as the supermatrix (de Queiroz and Gatesy 2006) and supertree (Bininda-Emonds 2004b) approach. The difference between these two methods lies in the level at which information is combined. In the supermatrix approach, a data matrix is constructed from multiple concatenated genes which is analysed to produce a tree. In the supertree approach, individual gene trees are built and then summarised to give a consensus tree. If all gene trees have the same terminal taxa, standard consensus methods such as Majority Rule (MR) can be used, while if the gene trees share partially-overlapping sets of terminal taxa, supertree methods must be used (Chen *et al.* 2006; Eulenstein *et al.* 2004; Wilkinson *et al.* 2005). The arguments in favour of the supertree approach are mainly based on perceived shortcomings of phylogenetic methods when applied to supermatrices. If the chosen tree reconstruction method fails to take into account differences in evolution between genes, then the method's

assumptions will be violated and the supermatrix approach will be inaccurate. This issue does not arise in the supertree approach since genes are analysed separately and models can be tailored to individual genes. Similarly, if the tree reconstruction method fails to correctly deal with missing data, the supertree approach might be considered superior, since taxa can only be included in an analysis when they have sequence data available for the gene in question. These arguments are rendered unconvincing in the face of modern phylogenetic methods that use likelihood and complex models, which can allow for both differences between genes and the presence of missing data. Since the final tree in supertree methods is derived from gene trees rather than a character matrix, it is an indirectly estimated tree and is constructed using only a subset of the data available (Bininda-Emonds 2004a; Gatesy, Baker and Hayashi 2004). Analysis of multiple genes using a supermatrix has been shown to yield support for clades that are not supported under analysis of any of the genes individually and would therefore not be recovered in a supertree (so-called 'hidden support'; de Queiroz and Gatesy 2006). While supertrees are the only viable approach in some circumstances (for instance, building phylogenies from trees where the data matrix is unobtainable), the supermatrix approach using modern phylogenetic methods has been very successful and is likely to be the best choice for multigene phylogenetics where the source data are available.

1.3 Obstacles to multigene phylogenetics

The problems associated with large-scale phylogenetic analyses can be divided into

those of data collection, data analysis and interpretation of results. Data collection refers to the process of turning raw sequence data into well-organised collections of sequences for each gene of interest. Data analysis refers to the methods used to draw phylogenetic conclusions from these sequences and typically consists of several stages. The sequences chosen for analysis are first aligned to produce a data matrix in which orthologous characters occupy the same column. This data matrix is then used as the input to a tree reconstruction program which generates a phylogenetic tree, normally with support values. Finally, the tree is interpreted to draw phylogenetic conclusions. Each of these steps has additional considerations when carried out in a multiple-gene, public sequence data framework. The TaxMan software package, described in detail in Chapter 2, implements possible solutions to the specific problems of orthology assignment (Section 1.3.1), choosing sequences for analysis (Section 1.3.2), multiple sequence alignment (Section 1.3.3) and inclusion of EST datasets (Section 1.4).

1.3.1 Orthology assignment

The problems of data collection are mainly a consequence of the fact that sequencing projects, and sequence databases, have generally not been designed with phylogenetics in mind. A major problem is the assignment of orthology. Orthologous sequences share their most recent common ancestor by virtue of a speciation event, in contrast with paralogues, which are related by gene duplication (reviewed in Koonin [2005]). Since an absolute requirement of phylogenetic analysis of taxa is that orthologous sequences are compared, it is essential when assembling a multigene dataset to ensure that all sequences for a given gene are orthologues. This problem takes different forms

depending on whether sequences are identified by looking at the sequence annotation, or at the sequence itself.

In the former approach, sequences with identical annotation are normally assumed to be orthologues. This approach is hampered by the fact that synonymy is rife among genes; researchers working on different organisms commonly use different names to refer to orthologous genes. Well-studied genes often have several synonyms (for example, the mitochondrial gene Cytochrome Oxidase 1 is variously annotated as COI, coxi, coxI, COX1, CO1, COXI, etc) and misannotation of sequences in public databases is not unknown.

In the latter approach, the sequences themselves are examined in order to assign orthology. In a fully *a priori* approach, sequences are clustered into orthology groups based on some measure of similarity. *A priori* orthology assignment is a weighty problem and methods for carrying it out are in their infancy. Multiple similarity metrics and clustering methods of varying degrees of automation and sophistication have been developed (Alexeyenko *et al.* 2006; Chiu *et al.* 2006; Dufayard *et al.* 2005; Li, Stoeckert and Roos 2003; Tatusov *et al.* 2003). This approach holds great promise for making use of previously under-utilised sequence data, and has been used successfully in some studies. However, there may be large differences in the orthology groups predicted by different methods and with different parameters, and a large amount of manual curation is currently necessary. A variant on this technique is to mine sequence databases for sequences with similarity to known genes. This approach is more well-developed, and software is available to do so using a variety of criteria (sequence similarity [Altschul *et al.* 1997]; Hidden Markov Models [Eddy 2001]), but

several issues limit the usefulness of this approach. The degree of similarity required to assign a sequence to a given gene is still somewhat arbitrary (although cross-validation techniques can offer a clue as to appropriate values). Another issue concerns the evolutionary distance between the sequences in the "known" gene set and those in the sequences to be evaluated. If the known gene set does not contain sequences representing the full taxonomic coverage of the sequences to be evaluated, similarity may be missed in a given candidate sequence when the most similar sequence in the known gene set is very distantly related. A third issue is that of gene families: several different genes from a given species may show nearly equal degrees of similarity to the known gene set. This is likely to be the case, for instance, where a gene has undergone a lineage-specific duplication (in this case, each instance is a co-orthologue of the known gene [Koonin 2005]). Determining which is the correct sequence to use for a species may require manual intervention. The issues outlined above mean that the similarity-based approach to sequence gathering is most successful for conserved, well-studied genes that are well-represented among annotated sequences (ensuring good taxonomic coverage in the known gene set) and which are unlikely to have paralogs in the species under study – the same qualities, in fact that have traditionally been required of genes used for deep phylogeny.

1.3.2 Choosing sequences, genes and taxa

A related problem is that of choosing, for a given gene in a given species, which of the available sequences to use for analysis. In many cases, mining public sequence databases will yield multiple sequences corresponding to the same gene in the same

species, and a choice must be made as to how they are to be used. Including partial sequences such as expressed sequence tags (ESTs) in a dataset raises particular issues that are dealt with in detail below.

The extremely large number of sequences in public sequence databases is the motivating factor for trying to solve the problems outlined above. However, the scale of sequence data is itself problematic in some ways, two of which I mention here. The GenBank sequence database contains around 80 million records for eukaryotic organisms and each record can contain multiple features of interest. Clearly, dealing with such a volume of information raises issues relating to data storage space and processing time. Accompanying this issue is the problem of updates and data freezes. A phylogenetic project that uses public sequence data will probably see new data become available over its lifetime. Bioinformatic solutions to the problems described above must take this into account and make provisions for new data to be incorporated into an analysis. A second consequence of working with very large data sources is that once sequences have been gathered and categorised, there are likely to be many more genes and species represented than can be included in an analysis. The computational complexity of an analysis increases rapidly with the number of characters and extremely rapidly with the number of sequences involved; for this reason it is usually better to include a set of genes and taxa that will answer the particular phylogenetic problem the researcher is interested in than to use all available data. The challenge is to balance the conflicting demands of good taxonomic sampling (requiring many species), a large amount of phylogenetic signal (requiring many characters) and the limitations of computing power. A confounding factor is the likely presence of

missing data in the dataset (i.e. the partial or complete unavailability of genes for some taxa). Techniques exist for selecting optimal sets of sequences from incomplete datasets but are not well-developed (Yan, Burleigh and Eulenstein 2005). The problem is made more difficult by the uneven distribution of sequence data over genes and species – model organisms and well-studied genes are likely to have far more sequence data than 'neglected' taxa and poorly-studied genes (Driskell *et al.* 2004). Additionally, the same dataset may be used to address different phylogenetic questions, each of which may well have different requirements.

1.3.3 Multiple Sequence Alignment

The essence of the alignment problem is simply stated: given a set of sequences that, since their divergence from a common ancestor, have undergone multiple substitution, insertion and deletion events, determine where in each sequence gaps should be inserted so that homologous characters occupy the same position (reviewed in Wallace, Blackshields and Higgins 2005). A multiple sequence alignment is a hypothesis, positing homologous relationships between characters. Since the computational time required to identify the globally optimal alignment increases exponentially with the number of sequences to be aligned, many heuristic methods for multiple sequence alignment have been developed (Edgar 2004; Higgins, Bleasby and Fuchs 1992; Lee, Grasso and Sharlow 2002). The best method to use for any given set of sequences is dependent upon the number and length of the sequences, the degree of similarity between them and the computational time available to solve the problem. In

general, alignment is more difficult for larger numbers of sequences and for sequences of greater length. The difficulty also increases as sequences become less similar, since orthology becomes more difficult to detect due to complex patterns of insertion, deletion and substitution. For deep multigene phylogenetics, sequence sets are likely to be both large (owing to the large amount of data) and highly diverged (owing to the large evolutionary distance between taxa), making the alignment process difficult. In addition, studies of this type are likely to include sequences for the same gene of varying length. Not only does sequence length vary naturally between taxa, but some sequences may be partial. This added difficulty calls for alignment methods that can cope with (1) large numbers of sequences, (2) a high degree of divergence and (3) the presence of partial sequences and sequences with varying length.

1.3.4 Evolutionary models and phylogenetic methods

The most advanced modern methods of tree reconstruction (Maximum Likelihood and Bayesian Inference) are likelihood-based. Both methods use an evolutionary model to derive a likelihood score (the probability of the data matrix given a tree topology and model parameters) for a given tree. They differ in the way this information is used. Maximum Likelihood methods attempt to find the tree and model parameters with the highest likelihood score, discarding all others. Bayesian Inferences takes trees and parameters with suboptimal scores into account to give an output that averages over uncertainty in the data. Because Bayesian analysis produces a large sample of trees, node support (posterior probabilities) can be estimated simply by counting the frequency of each node in the sampled trees. In contrast, bootstrapping (Felsenstein

1985) is normally used to derive node support values under Maximum Likelihood. Philosophical and theoretical arguments can be used to favour one or other method, but such discussions are outside the scope of this thesis (Huelsenbeck and Rannala 2004; Sullivan and Swofford 2001; Svennblad *et al.* 2006).

Key to both methods are evolutionary models. An evolutionary model describes the probability of character change along a branch and can be used to evaluate the likelihood of a particular topology, the key step in likelihood methods. Traditional models describe the relative proportions of different character states (in molecular terms, nucleotides or amino acid residues), the rates of transitions between states, and the variation in rates among sites. Likelihood methods have largely used the GTR family of models with rate variation approximated by gamma categories and proportions of invariant sites. Recently, evolutionary models have begun to include a greater degree of realism and complexity by dividing characters up into partitions and allowing different partitions to have different patterns of evolution (Nylander *et al.* 2004). Data matrices are commonly divided up into different genes, into different codon positions, or into stem and loop structures in RNA genes. Any data matrix where there is an *a priori* expectation for different patterns of evolution in some set of characters is a candidate for a partitioned model. New methods even allow investigation of non-obvious partitions in a data matrix (Pagel and Meade 2004). In multigene studies, the argument for data partitioning is very convincing, as models that do not take partitions into account have been shown to describe the data suboptimally and to negatively influence phylogenetic reconstruction (Brandley, Schmitz and Reeder 2005; Pupko *et al.* 2002; Chapter 3). Not all phylogenetic software supports

partitioned models, so researchers carrying out multigene studies will find themselves limited in the methods they can use. Model choice can be driven by knowledge about the molecular evolutionary processes or by formalised statistical model testing. Additionally, simulation studies have been used to investigate the effects of model misspecification (Lemmon and Moriarty 2004).

Likelihood methods have another property that is crucial for analysis of public-sequence-derived datasets; they deal correctly with the presence of missing data by averaging the likelihood score over all possible character states. Other phylogenetic methods, such as distance methods, are not so robust to missing data and may not be suitable for analysis of this type of data. Even within the likelihood methods, resampling techniques like bootstrapping can cause problems when applied to datasets with missing data and care is needed. Simulations have shown that, under likelihood methods, missing data is not necessarily a barrier to phylogenetic accuracy (Kearney 2002; Wiens 2003; Wiens 2006).

Computational complexity, as at every stage of large-scale phylogenetics, is an issue in the choice of phylogenetic methods. As we have seen, large data matrices are the norm in this type of study and so only methods and software that can analyse large amounts of data in a reasonable time, including the generation of support values, will be suitable. Taking into account the need for (1) support for complex, partitioned models, (2) robustness to missing data and (3) rapid analysis of large datasets, Bayesian Inference emerges as the best phylogenetic method currently available for multigene deep phylogeny. The average researcher has to work within the limitations of not just the method but the particular implementation offered in a given piece of

software. Bayesian Inference is well served by several software packages (Drummond and Rambaut 2003; Pagel and Meade 2004; Ronquist and Huelsenbeck 2003) that fulfill the requirements outlined above.

1.3.5 Sources of bias

As outlined above (Section 1.1), single genes can fail to resolve relationships when the number of characters is small. Multigene phylogenetics circumvents this problem by including more characters in a data matrix, increasing the amount of phylogenetic signal available. The multigene approach marks a shift away from lack of phylogenetic signal and towards systematic bias as the factor limiting phylogenetic accuracy. With a single-gene data matrix a fixed number of characters is available for analysis and, regardless of the methods used, if there is no phylogenetic signal supporting a particular relationship it will not be recovered. For an extreme example, imagine a phylogenetic study which includes a monophyletic group defined by a very short branch. A short gene with a low evolutionary rate may, by chance, have not undergone any substitutions along that branch, in which case phylogenetic analysis of that gene will never robustly recover the monophyletic relationship. If, however, we could add genes to the data matrix, increasing the number of characters analysed, characters that had undergone substitutions along that branch would stochastically be added to the data matrix and support for the monophyly of the group would steadily increase. In this scenario, phylogenetic accuracy is being limited by the amount of signal present in the dataset.

By contrast, imagine a second phylogenetic study which includes two distantly related species with abnormally rapid evolutionary rates. Analysis of any single gene results in recovery of an erroneous sister taxon relationship between these two species due to a long branch attraction artefact. Crucially, in this scenario, adding additional characters will not help to resolve the phylogeny correctly. Instead, it will merely increase the robustness of the incorrect result, because the bias is systematic; i.e. it affects all genes equally. This effect, called inconsistency, was first described by Felsenstein (1978) for the case of long branch attraction, but other sources of systematic bias have since been described including base composition (Foster and Hickey 1999), mitochondrial strand bias (Hassanin, Leger and Deutsch 2005; Chapter 3) and heterotachy (Gadagkar and Kumar 2005; Philippe *et al.* 2005). Using multiple genes can increase phylogenetic accuracy in cases where it was previously limited by lack of phylogenetic signal, as in the first scenario, but not in cases where is is limited by systematic bias, as in the second scenario. Sources of systematic bias must be investigated if multigene studies are to be found convincing.

1.4 Using partial genomes

Of the ~80 million eukaryote records in GenBank, ~40 million are expressed sequence tags (ESTs [Boguski, Lowe and Tolstoshev 1993]). ESTs are short (~600 bp) single-pass reads of randomly cloned mRNAs and represent transcribed DNA sequences. EST projects are a cheap and technically straightforward way to sample the transcriptome of an organism of interest and represent a valuable source of sequence

information for phylogenetics (Hughes *et al.* 2006; de la Torre *et al.* 2006). Because EST sequencing requires only a relatively small investment of resources, it is an excellent way to fill in the taxonomic gaps in sequence coverage, and can therefore be used to obtain sequence data from organisms that do not warrant greater sequencing effort. ESTs have characteristics that pose particular challenges for phylogenetics and which any multigene study must take into account. Because ESTs are short relative to most genes, any individual EST sequence is likely to represent only part of the coding sequence. The single-read nature of EST sequencing makes the sequences error-prone, with an estimated error rate of 1%. Because mRNA molecules are randomly sampled to generate EST libraries, unless the libraries are normalised genes will be present in proportion to the abundance of their mRNA in the cells from which the library was generated. This leads to redundancy within most EST projects, with highly-expressed genes represented by multiple EST sequences. A collection of ESTs is usually processed to remove redundancy, increase the length of gene sequences, and improve the accuracy of the sequences (Parkinson *et al.* 2004a; Pertea *et al.* 2003). The first step in EST processing is to cluster the sequences into gene objects on the basis of similarity. A consensus sequence is then built for each cluster which incorporates the information from each individual sequence. Because the ESTs are normally partially overlapping, the consensus sequence is longer than any of the individual sequences. Additionally, sequencing errors in one EST sequence can be overruled by correctly-called bases in other EST sequences that overlap. The consensus sequence is more suitable for phylogenetic analysis since it contains a greater amount of information and is more accurate. Typically, the raw ESTs generated for an EST project are deposited

in GenBank while the processed sequences are made available via a dedicated website (e.g. Parkinson *et al.* 2004b). This allows interested researchers to carry out their own processing if necessary.

To use ESTs for multigene phylogenetics requires that the identity of each consensus sequence is determined so it can be assigned to the appropriate orthology group. BLAST (Altschul *et al.* 1997) similarity to known sequences is often used as an identification criterion as described in Section 1.3.1 and is further discussed in the chapter describing TaxMan (Section 2.5.2).

1.5 Thesis summary

1.5.1 Bioinformatics

The sections above lay out the motivation for attempting to use public sequence data to carry out large scale, multigene deep phylogenetics, and briefly cover the obstacles to doing so. In practical terms, any attempt to carry out such a study must rely heavily on bioinformatics techniques. Bioinformatics refers to the use of computers to store and manage biological data, and of software techniques to analyse it. Its rise as a field has been largely driven by the large volumes of sequence data that form the subject of earlier sections. At its simplest, bioinformatics can mean using Entrez (web reference 1) to select sequences to download from GenBank, or pasting a nucleotide sequence into an online BLAST server. Increasingly, however, biologists are taking up programming languages to write software specific to their needs. The concept of codifying biological knowledge and best-practice procedures is a powerful one, as it

allows much greater volumes of data to be processed than could be achieved manually. In most bioinformatics software the aim is to reduce manual data curation as far as possible while implementing sensible analyses. This approach has been very successfully applied to gene finding (Burge and Karlin 1997), EST processing (Wasmuth and Blaxter 2004), protein domain classification (Bateman *et al.* 2002), etc. In chapter 2 I describe the design and implementation of a software package, TaxMan, which applies these bioinformatics principles to the problem of assembling a large dataset of aligned gene sequences from public sequence data. I give an overview of current software solutions, outline the design assumptions and discuss the most important features of TaxMan, before summarising some benchmarking results. The software is freely available for download and the user guide, containing practical guidelines for working with large sequence datasets, is provided as Appendix 1. A manuscript describing TaxMan is in press at BMC Bioinformatics, and is provided as Appendix 2.

1.5.2 The phylum Chelicerata

One potential application of multigene phylogenetics from public sequence data is to quickly place new data in context. New sequence data (particularly EST data) is typically annotated by comparison to existing sequences. Similarly, existing libraries of protein motifs or protein families can be used to gain an insight into the function of a newly sequenced gene. In the same way, when new sequence data is obtained for a species one can assemble orthologous sequences from related taxa and carry out phylogenetic analysis to determine the relationships of the species. In chapter 3 I

describe the application of this approach to a newly sequenced mitochondrial genome, that of *Mesobuthus gibbosus*, sequenced by collaborator Benjamin Gantenbein. I use TaxMan to assemble a dataset of mitochondrial genes from related chelicerate taxa and investigate the use of mitochondrial gene datasets to determine the phylogenetic position of scorpions within Chelicerata. The chapter includes discussions on several issues mentioned above – the skewed distribution of data across genes and species; the presence of systematic bias in multigene datasets, and the importance of model choice for phylogenetics. A manuscript describing this work is in press at Molecular Phylogenetics and Evolution, and is provided as Appendix 3.

1.5.3 The superphylum Lophotrochozoa

Another application of multigene phylogentics from public sequence data is to rapidly assemble datasets to investigate newly erected phylogenetic groups. The most striking recent example are the Ecdysozoa and Lophotrochozoa, two superphyla that together comprise Protostomia. Evidence for the monophyly of these two clades came from 18S ribosomal RNA data but evidence for (and against) them has subsequently been found in a number of datasets. The Ecdysozoa, comprising moulting animals, includes the well-studied Arthropoda and Nematoda and is well-represented in GenBank, with ~5.5 million records. The Lophotrochozoa, comprising phyla with a trochophore larva or a lophophore feeding structure, is much less well-represented in GenBank, with just ~0.8 million records. Consequently, many more large-scale molecular phylogenetic studies have been carried out on ecdysozoan taxa (especially arthropods) than on lophotrochozoan taxa. In chapter 4 I describe the use of TaxMan to assemble a dataset

of mitochondrial and nuclear protein-coding and ribosomal RNA genes for the lophotrochozoa and use the dataset to explore issues in lophotrochozoan phylogeny.

2 Taxman

2.1 Abstract

Phylogenetic analysis of large, multiple-gene datasets, assembled from public sequence databases, is rapidly becoming a popular way to approach difficult phylogenetic problems. Supermatrices (concatenated multiple sequence alignments of multiple genes) can yield more phylogenetic signal than individual genes. However, manually assembling such datasets for a large taxonomic group is time-consuming and error-prone. Additionally, sequence curation, alignment and phylogenetic analysis are made particularly difficult by the potential for a given gene in a given species to be unrepresented, or to be represented by multiple or partial sequences. I have developed a software package, TaxMan, that largely automates the processes of sequence acquisition, consensus building, alignment and taxon selection to facilitate this type of phylogenetic study.

TaxMan uses freely available tools to allow rapid assembly, storage and analysis of large, aligned DNA and protein sequence datasets for user-defined sets of species and genes. The user provides GenBank format files and a list of gene names and synonyms for the loci to analyse. Sequences are extracted from the GenBank files on the basis of annotation and sequence similarity. Consensus sequences are built automatically. Alignment is carried out (where possible, at the protein level) and aligned sequences are stored in a database. TaxMan can automatically determine the best subset of taxa to examine phylogeny at a given taxonomic level. By using the

stored aligned sequences, large concatenated multiple sequence alignments can be generated rapidly for a subset and output in analysis-ready file formats. Trees resulting from phylogenetic analysis can be stored and compared with a reference taxonomy.

TaxMan allows rapid automated assembly of a multigene datasets of aligned sequences for large taxonomic groups. By extracting sequences on the basis of both annotation and BLAST similarity, it ensures that all available sequence data can be brought to bear on a phylogenetic problem, but remains fast enough to cope with many thousands of records. By automatically assisting in the selection of the best subset of taxa to address a particular phylogenetic problem, TaxMan greatly speeds up the process of generating multiple sequence alignments for phylogenetic analysis.

Some of the material in this chapter has been written as a paper, currently in press at BMC Bioinformatics. Martin Jones wrote the software and manuscript. Mark Blaxter assisted with software design and testing and supervised the project.

2.2 Introduction & Background

2.2.1 Motivation

Traditionally, phylogenetic analyses of large taxonomic groups have been carried out by sequencing selected single genes for particular chosen taxa. In such analyses, both genes and taxa are chosen for their suitability for phylogenetics. Genes are selected which are thought to display the characteristics required of good phylogenetic markers. Such characteristics include an appropriate rate of evolution for the group in question,

a reasonably constant rate of evolution across the taxa sampled, ease of orthology assignment, ease of alignment and ease of sequencing. Taxa are selected to fulfil two criteria: broad taxonomic sampling across the study group and inclusion of representatives from any clades that are of particular importance in the evolutionary hypotheses to be tested. This approach to phylogenetics has been extremely successful in many cases, resolving relationships where analysis of morphological characters has failed to give a clear answer. However, such studies usually involve few genes and hence a limited amount of phylogenetic information, and have thus been unable to solve some difficult phylogenetic problems, particularly those involving ancient relationships.

With the growth of public sequence databases (e.g. GenBank [Benson *et al.* 2006]) a different approach to phylogenetics has become possible. In this new approach, existing gene sequences are obtained from public databases rather than being produced specifically for the study. This leads to a number of differences relative to the classical type of phylogenetics study that greatly affect the research strategy. Instead of targeting genes and taxa on the basis of their phylogenetic utility, researchers are limited to sequence data that has already been produced, possibly for different types of analysis and consequently according to different criteria. For most large groups, the distribution of sequence data across genes and species will be highly skewed (Driskell *et al.* 2004; Sanderson and Driskell 2003). Generally, a small number of genes, having been intensively studied, will be available for a large number of taxa while the majority of genes will be available for only a small number of taxa. Likewise, in a large taxonomic group a few species, particularly model organisms, will have a great

deal of sequence data (and hence many genes available) while the majority, often referred to as neglected taxa, will have very few sequences. Interestingly, this phenomenon can be seen at different taxonomic levels – within a phylum, one class will be over-represented (e.g. Insecta within Arthropoda) and within a family, one genus will be over-represented (e.g. *Drosophila* within Drosophilidae).

The choice facing the phylogeneticist when assembling a large sequence dataset is to balance the number of genes, number of taxa and proportion of missing data (Rokas and Carroll 2005). If a complete or nearly complete dataset is required, either the number of genes or the number of taxa included in the analysis must be small. At one extreme lie studies involving a few well-studied species with fully sequenced genomes (Rokas *et al.* 2003). Using whole genomes for phylogenetics necessarily results in a sparse sampling of taxa, and many important groups may be represented by only a single species. Species for which whole genome sequences are available have generally been chosen for criteria (model organism, disease vector) that may make them poor exemplars of their respective groups. Additionally, artefacts such as long branch attraction are more likely in an analysis with sparse taxon sampling. The other extreme is represented by studies on a single gene that include many taxa. Single genes may be misleading or unsuitable for phylogenetic analysis for several reasons. A gene that has undergone accelerated evolution in one lineage (or, in more general terms, whose pattern of evolution has differed between lineages) would not be suitable for phylogenetic analysis. This issue is particularly acute in likelihood methods of phylogenetics (including Bayesian reconstruction and Maximum Likelihood) where a single model of evolution is usually assumed to apply throughout a tree (Philippe *et al.*

2005). Additionally, a given gene's sequence may be too well- or poorly- conserved for phylogenetic reconstruction at a given taxonomic level.

In many cases, particularly those involving deep phylogeny, it is inadvisable to rely on datasets including few genes or few taxa. Phylogenetic accuracy in difficult cases has been shown to depend both on a large number of characters (providing phylogenetic signal) and on adequate taxonomic sampling (ensuring good representation of a clade's diversity). An alternative strategy to the two extremes outlined above is to include multiple genes and a broad sampling of taxa, and in addition allow for the presence of missing data while striving to minimise its impact (Driskell *et al.* 2004; Sanderson *et al.* 2003; Sanderson and Driskell 2003; Yan, Burleigh and Eulenstein 2005). This approach allows the use of a larger volume of sequence data, including genome-scale data and partial genomes derived from expressed sequence tags (ESTs). ESTs are a powerful tool for gene sampling from neglected taxa, and can greatly increase the taxonomic sampling in a dataset. Such a strategy requires that the availability of sequence data for each gene must be assessed for each species in the taxonomic group of interest. This can be time consuming, particularly for taxonomic groups with large numbers of species.

Similar challenges have been faced in many areas of research where rapidly increasing volumes of data make manual analysis difficult. Such areas have been well served by using bioinformatics to facilitate an automated approach. Tasks such as gene functional annotation (Martin, Berriman and Barton 2004), EST processing (Parkinson *et al.* 2004a) and genome comparisons (Goodstadt and Ponting 2006) have been rendered tractable by high-throughput, automated pipelines. A similar strategy is of

benefit to phylogenetic dataset assembly. An automated approach to dataset assembly offers several advantages. The time required to assemble a large sequence dataset is greatly reduced, allowing more emphasis to be placed on quality assessment and analysis. Automated sequence acquisition can be more thorough than manual curation and can cope with volumes of data that would be unfeasible to sort manually. Additionally, a standardised dataset assembly procedure allows for rapid rebuilding of datasets in response to new sequence data becoming available, meaning that it is easier to keep a sequence dataset current. Depending on the implementation, heterogeneous data (raw sequences, consensus sequences, protein translations etc.) can be stored in a relational manner, further reducing the burden of organisation and facilitating record-keeping. Such an approach also takes care of issues of data provenance, ensuring that the results of downstream analysis can always be traced back to the raw sequence data. Finally, the process of dataset assembly can be valuable, even if phylogenetic analysis is not the intended outcome, as a means to provide an overview of the distribution of sequence data with a large taxonomic group. The software package TaxMan is intended to realise all of the above benefits.

2.2.2 Related software

The problem that TaxMan solves is not new; the potential of large scale phylogenetics, and the inherent difficulties, have been apparent for some time. Several workers have considered a bioinformatics-based solution to these problems. Although there is no software currently available with the same features as TaxMan, multiple packages

exist that share at least some of their aims with TaxMan, and are thus relevant for discussion. Table 2.1 gives a comparison of some key features of the tools discussed below.

Package name	multiple genes	sequence extraction	consensus building	stores aligned sequences	subsets	stores taxonomy	carries out alignment	carries out phylogenetic reconstruction	tree editing	alignment editing	sequence clustering
TaxMan	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*yes*	*no*	*no*	*no*	*no*
ARB	no	no	yes	yes	yes	yes	yes	yes	yes	yes	no
Phylota	no	no	no	no	yes	no	no	no	no	no	yes
HAL	yes	yes	no	no	no	no	yes	yes	no	no	no
TreeBlaster	no	yes	no	yes	no	no	yes	yes	no	no	no
AmiGA	yes	no	no	no	yes	yes	yes	no	no	no	no
MUST	yes	no	no	yes	yes	yes	no	no	no	yes	no

Table 2.1: Comparison of some key features in TaxMan and other software

ARB

The ARB project (Ludwig *et al.* 2004) is a software tool designed for analysis of ribosomal RNA data (though it can also be used for protein-coding genes). It includes a database of stored aligned sequences and a graphical user interface through which analyses can be carried out. Phylogenetic analysis can be carried out on subsets of sequences from within the ARB environment, and sequences can be organised into taxonomic schemes. ARB can align sequences using CLUSTALV (Higgins, Bleasby and Fuchs 1992) and can create consensus sequences from user-specified sets of

sequences, although it has no concept of sequence type hierarchy. It cannot cope with multiple genes or extract relevant sequences from GenBank records, and uses a straightforward alignment strategy where all sequences to be aligned are simply passed to the CLUSTALV program. ARB includes a number of tools not directly related to phylogenetic studies, such as primer and probe design, and secondary structure editing. In general, ARB is a good choice of program for managing a large dataset of intensely-studied sequences (such as ribosomal RNA genes) but is unsuitable for large-scale phylogenetics studies involving multiple genes.

Phylota

The unpublished Phylota project (web reference 2) aims to develop biological tools for extracting phylogenetic information from sequence databases and using it to assemble phylogenetic trees. Currently available software includes tools for database clustering (of use in defining orthologous relationships between genes) and for visualising the clusters. Tools are also available for defining maximally complete subsets of sequences (quasi-bicliques), although there is no facility to produce alignments. At this time, the Phylota project is not capable of extracting or storing sequences, or of building or aligning consensus sequences. The tools made available by the Phylota project could be very useful for incorporation in future work on TaxMan.

HAL

The unpublished HAL package (web reference 3), part of the fungal Tree of Life project, is a set of scripts that constitute a pipeline for identifying orthologues from proteomes and producing concatenated multiple sequence alignments. For a given set of input proteomes, HAL carries out all-versus-all BLASTP (Altschul *et al.* 1997)

searches and uses the results as the input to TribeMCL (Enright, Van Dongen and Ouzounis 2002), a Markov Flow Clustering program. HAL carries out TribeMCL analysis with a range of parameter sets to identify putative orthologue groups. Sequences for each orthologue group are aligned using CLUSTALW and an evolutionary model is assigned to each using ProtTest (Abascal, Zardoya and Posada 2005). Finally, alignments are concatenated to form a supermatrix (referred to as a 'Super Alignment' in the HAL documentation) which is analysed using PAUP (Swofford 2006), PHYLIP (Felsenstein 2005) and PHYML (Guindon and Gascuel 2003). HAL contains many useful ideas for orthology assignment, a task that is outwith TaxMan's design brief. However, it cannot identify sequences based on annotation, or extract sequences from GenBank format files, and has no consensus building tools. It cannot use subsets of data and sequences cannot be organised taxonomically. HAL would be a good choice for researchers working with fully-sequenced genomes and proteomes, who wanted to use anonymous orthologue groups for phylogenetic analysis. The ideas used in HAL would be very useful in further development of TaxMan.

TreeBlaster

The unpublished TreeBlaster package (web reference 4) from the Protist EST Program (PEP) project (web reference 5) is a web-based application for extracting sequences from public databases using BLAST similarity and aligning them. Sequences showing significant sequence similarity to a 'seed' sequence can be added to a sequence file. The process it iterative; added sequences may be used as the input for further rounds of sequence addition. Once all desired sequences have been added, they are aligned using

CLUSTALW (Thompson, Higgins and Gibson 1994) and then are submitted to a phylogenetics pipeline comprising PUZZLE (Schmidt *et al.* 2002), WEIGHBOUR (Bruno, Socci and Halpern 2000) and CONSENSE (Felsenstein 2005) to build phylogenetic trees. TreeBlaster implements a subset of the functionality offered by TaxMan, in that it (1) extracts sequences from public databases based on sequence similarity and (2) carries out multiple sequence alignment, but lacks many of TaxMan's capabilites.

AMiGA

AMiGA, the Arthropodan Mitochondrial Genomes Accessible database (Feijao *et al.* 2006), is not a software package but rather a web-accessible database of mitochondrial genes derived from whole mitochondrial genomes. Nevertheless, it shares several key features with TaxMan, and is thus a suitable subject for discussion. The database holds sequences for protein-coding, ribosomal RNA and transfer RNA genes for fully-sequenced arthropod mitochondrial genomes. One of the uses of AMiGA is the generation of concatenated multiple sequence alignments and it is in this respect that AMiGA shares similarities with TaxMan. When producing alignments, subsets of taxa can be chosen in a hierarchical taxonomic framework – higher taxonomic groups (e.g. orders) can be selected in addition to individual species. Subsets of genes can also be chosen for analysis. This functionality is similar in intent to TaxMan's slices (see Section 2.3.6), with the exception that AMiGA cannot generate subsets automatically. Additionally, AMiGA can generate supermatrices containing multiple concatenated genes for the taxa specified. A major difference between TaxMan and AMiGA is the way in which sequences are stored; AMiGA, in contrast to TaxMan, stores unaligned

sequences and performs multiple sequence alignment at the time the sequences are requested by the user, making it much slower than TaxMan for some queries. Though a user cannot extract or add sequence information to the AMiGA database, it does include some of the same ideas as TaxMan and is an interesting example of a useful web-based database with explicitly phylogenetic intentions. AMiGA has a rich set of genome comparison tools that are not relevant here, since their functionality does not overlap with that of TaxMan.

MUST

MUST (Management Utilities for Sequences and Trees; Philippe 2003) is a collection of programs for storing sequences and carrying out phylogenetic analysis. It imports sequences in a number of formats (GenBank, EMBL) and uses external programs to align them. The alignment can be manually edited. It stores sequences in a phylogenetic context and allows the user to select subsets of taxa for analysis, which is carried out by external programs. While remarkably forward-looking for its time, MUST is no longer under development. It does not support the most recent phylogenetic and alignment software and cannot extract sequences from large databases.

2.2.3 Features of TaxMan

TaxMan allows the user to mine public sequence data in order to rapidly assemble a large dataset of aligned genes for use in phylogenetic analysis. It uses annotation present in the GenBank sequence database to identify sequences of interest in GenBank records. It can also identify sequences on the basis of similarity to known

genes, allowing EST sequence data to be included. Once sequence data is collected, TaxMan builds and aligns consensus sequences, storing the aligned sequences for rapid retrieval. TaxMan can use this stored data to produce analysis-ready alignment files containing mutiple genes for a given subset of taxa. Because TaxMan also stores the NCBI taxonomy (web reference 6), it can automatically select the best set of taxa to address phylogenetic questions at a given taxonomic level. Phylogenetic analysis of alignments result in trees which can be stored and retrieved by TaxMan; trees can also be compared to the NCBI taxonomy.

TaxMan maximises the amount of phylogenetic information available for analysis by including all types of sequence data (including unannotated EST data) and ensuring that the highest-quality sequence is used for consensus-building. TaxMan supports phylogenetic best practice by allocating separate partitions to genes and codon positions in the alignment files it produces. TaxMan is fast; tens of thousands of GenBank records can be easily processed in a single working day, allowing a researcher to assemble much larger datasets that would be possible manually. Below I describe the design and implementation of TaxMan.

2.3 Design

2.3.1 Assumptions

TaxMan follows several general dataset assembly strategies that differ from those employed in classical phylogenetic studies. These general principles dictate the

approach taken in the design of TaxMan.

Defined gene and taxon set

TaxMan is designed within a series of assumptions about the type of data that the user will be working with. Specifically, it assumes that the user has assembled (1) a list of gene names and synonyms, and (2) a collection of GenBank records for the taxa of interest. A further assumption is that the genes selected are single copy orthologues; i.e. that any GenBank sequence accurately annotated with one of the gene name synonyms is orthologous to all other sequences so labelled. Using sequences in GenBank format allows TaxMan to assume that all sequences can be assigned to a species that has been allocated an NCBI taxonomic identifier (taxid hereafter) and that the species appears in the NCBI taxonomy. Based on these assumptions, TaxMan assembles a dataset consisting of a single aligned consensus gene sequence for each gene in each species where sequence data is available.

Sequence gathering

In contrast to traditional studies, where sequences are obtained for a predefined set of taxa, TaxMan's approach is to begin by examining all available sequences for the selected genes in the taxonomic group under consideration. This ensures that as much sequence data as possible is available for phylogenetic analysis. Records are obtained by the user from Entrez (web reference 1) (see Section 2.3.3 for details). Depending on the size of the group and degree of sequencing effort directed towards it, this can be on the order of $10^5 – 10^6$ records. This approach is particularly valuable for the study of taxonomic groups where the distribution of sequence data is not known *a priori*, as choices regarding the set of genes and taxa to be used for phylogenetic analysis are

deferred.

Align once

TaxMan carries out a single alignment event for each selected gene and stores the aligned sequences (see Section 2.3.5 for details of the alignment process). This philosophy ensures alignments containing subsets of genes and taxa can be produced very quickly in a format suitable for phylogenetic analysis. TaxMan's design is based on the assumption that multiple subsets of the data will be produced and analysed in order to explore the phylogenetic signal in the dataset. On this assumption, the align-once strategy is most effective.

2.3.2 Overview diagram & strategy

Figure 2.1 gives an overview of the TaxMan processing pipeline. The first part of the pipeline is concerned with producing a dataset of aligned orthologous genes and is generally carried out only once for a given dataset. The second part of the pipeline is concerned with producing alignments containing subsets of the data, subjecting them to phylogenetic analysis, and storing and retrieving the resulting trees. It is normally carried out multiple times for a given dataset. There are three stages involved in the first part of the pipeline. Firstly, raw sequence data pertaining to the genes and taxa of interest must be gathered. Second, a consensus sequence must be built for each gene in each species. Finally, these consensus sequences must be aligned for each gene.

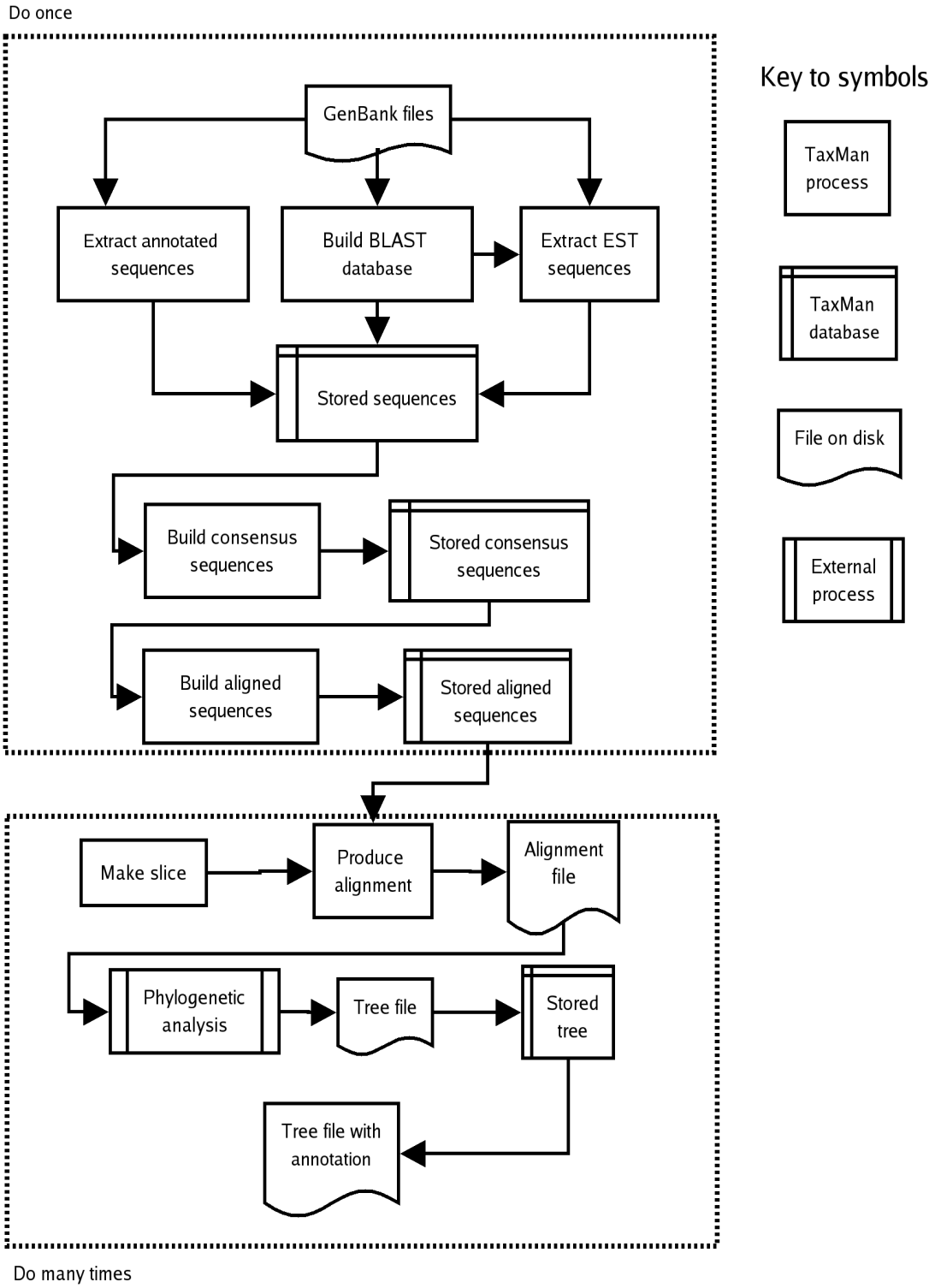


Figure 2.1: Overview of the TaxMan workflow

2.3.3 Sequence gathering

The two primary concerns when gathering raw sequence data for a large-scale, multigene phylogenetic study is that all available sequence data is used, and that orthology is correctly assigned. Failure to address the first concern will result in loss of potential phylogenetically informative data; failure to address the second could have much more severe effects on phylogenetic accuracy. TaxMan takes its raw sequence data in the form of GenBank format files, each of which can contain multiple GenBank records. The GenBank record format allows a DNA or protein sequence to be annotated with various pieces of metadata; in particular, it allows labelling of coding regions corresponding to genes (see Section 2.3.5). TaxMan uses this annotation to identify sequence data belonging to a particular gene. Because different researchers use different names to refer to the same gene, a well-studied gene of the sort likely to be useful for phylogenetic analysis will commonly have multiple synonyms. For this reason, the user must supply a list of synonyms for each gene that is to be included in the dataset. For each gene, the user gives a canonical name (by which the gene is to be referred to within TaxMan) and a list of synonyms that may possibly appear in the GenBank file annotation. Thus, for the gene ***Cytochrome Oxidase subunit 1*** we have the canonical name:

COX1

and the synonyms

COI, coxi, coxI, COX1, CO1, COXI, co1, coi, cox1, cytochrome oxidase subunit 1, coI, cytochrome oxidase

subunit I

For each gene, a genetic code is also specified. This allows genes from different organelles to be combined in a single dataset (e.g. nuclear and mitochondrial genes).

The different types of sequence

Because of TaxMan's approach to sequence acquisition, sequences fall into two categories – those that have been identified on the basis of annotation and those that have been (putatively) identified on the basis of sequence similarity. These two categories determine the level of confidence in the assignment and quality of the sequence and control how it is used during processing. Sequences identified by annotation (referred to hereafter as 'annotated sequences') are assumed to be more likely to have been correctly identified than those identified on the basis of BLAST (Altschul *et al.* 1997) similarity (referred to as 'screened sequences'). Additionally, annotated sequences are likely to be higher quality and more likely to be full length than screened sequences. This is due to the fact that many screened sequences come from EST projects. EST sequences are known to have a relatively high error rate due to the single-read nature of the sequencing, and often represent only partial messenger RNAs. Taken together, these properties define a hierarchy of sequence types. If we define a subclass of annotated sequences which come from fully sequenced mitochondrial or nuclear genomes, then we can state a progression in confidence of identification, sequence length and sequence quality from screened sequences, to annotated sequences, to annotated sequences that are derived from genomes.

Extracting sequences by annotation

To extract sequences based on GenBank format file annotation, TaxMan reads a list of gene names and synonyms from the TaxMan config file and looks for matching features annotated in a GenBank format file. Features are sub-sequences within a DNA sequence that have some associated metadata. If a feature is of the type 'gene', 'product' or 'RNA' then its 'gene' or 'product' tag is checked against the list of synonyms stored in the TaxMan database. If the annotation matches one of the synonyms then the sequence is extracted, tagged with the canonical gene name and stored in the TaxMan database. To avoid incorporating incorrectly annotated sequences into the database, TaxMan takes a conservative approach and only includes features whose annotation *exactly* matches one of the gene synonyms. Allowing partial matches would potentially increase the number of sequences extracted, but would risk silently including sequences with similar annotation that were not orthologous. It is particularly important to be conservative with sequence extraction at this early stage as errors here could propagate through later stages of analysis.

Extracting sequences by similarity

In addition to sequence acquisition by annotation, TaxMan can also extract sequences from GenBank files on the basis of BLAST (Altschul *et al.* 1997) sequence similarity to known sequences. This allows the user to make use of the large amount of unannotated sequence data (much of which is derived from EST projects) and any sequences annotated with unknown synonyms. In order to screen sequences for similarity, a BLAST database of known genes must be generated. TaxMan can do this automatically by extracting sequences from an annotated GenBank file, using the

process outlined above, optionally using a 'whitelist' of NCBI taxids specifying the species to be included in the BLAST database. The construction of a BLAST database in this fashion requires a trade-off between database size (and consequent search time) and taxonomic coverage. Including sequences from many species in the BLAST database can increase the taxonomic coverage, making it more likely that a relevant sequence will have a match from a closely related species in the database and hence increasing the number of screened sequences extracted. However, it will also increase the processing time required to search for screened sequences.

Assigning sequences to species, and species to higher groups

Sequences are allocated to species by using the NCBI taxid that is stored as part of the metadata for each record. When all sequences of interest have been extracted, species information is stored for all represented species. For each represented species, TaxMan stores the genus and species name, the NCBI taxid and the common name. The NCBI makes the structure of its taxonomy available by publishing a 'taxonomy dump' file which contains details of the relationships between species and higher taxonomic groups (web reference 7). TaxMan parses this file in order to automatically assign species to genus, family, order and class. For a given taxonomic level, TaxMan allocates species to groups at that level in a way that partitions the species into non-overlapping sets. For example, when assigning families:

- where possible, the node labelled with rank 'family' in the lineage for any given species is used as the family for that species.

- if the node labelled with rank 'family' in the lineage for a given species is a child of another node that has been used as a family it cannot be used (as this would result in overlapping family groups), and the next available higher-level node must be used.

Orders, classes and genera are allocated in a similar way. These criteria are necessary in order to ensure that, at any given level, the species stored by TaxMan can be partitioned into non-overlapping sets. It is a consequence of the incomplete nature of the NCBI taxonomy, in which many species have no node labelled 'family' in their lineage. Since the NCBI taxonomy is the only classification scheme which is used to organise sequences stored in GenBank, it is used by TaxMan, with appropriate warnings to the user. The genus, family, order and class allocations are used by TaxMan when building slices automatically (see Section 2.3.6)

2.3.4 Consensus building

By design, TaxMan deals only with predefined orthologous groups of sequences. When assembling a dataset, the following assumptions are made:

- All sequences annotated with one of the synonyms of a particular gene are orthologous
- A sequence showing highly significant BLAST similarity to a known gene is orthologous to it
- All loci listed in the config file are single-copy in all taxa under consideration
- All loci in all species can be represented by a single consensus sequence

Given these assumptions, it is necessary for each gene in each species to be

represented by a single consensus sequence. Using the hierarchy of sequence qualities described above, TaxMan follows a set of rules for deriving consensus sequences that satisfies two criteria; that consensus sequences be as accurate, and as complete, as possible (see Figure 2.2).

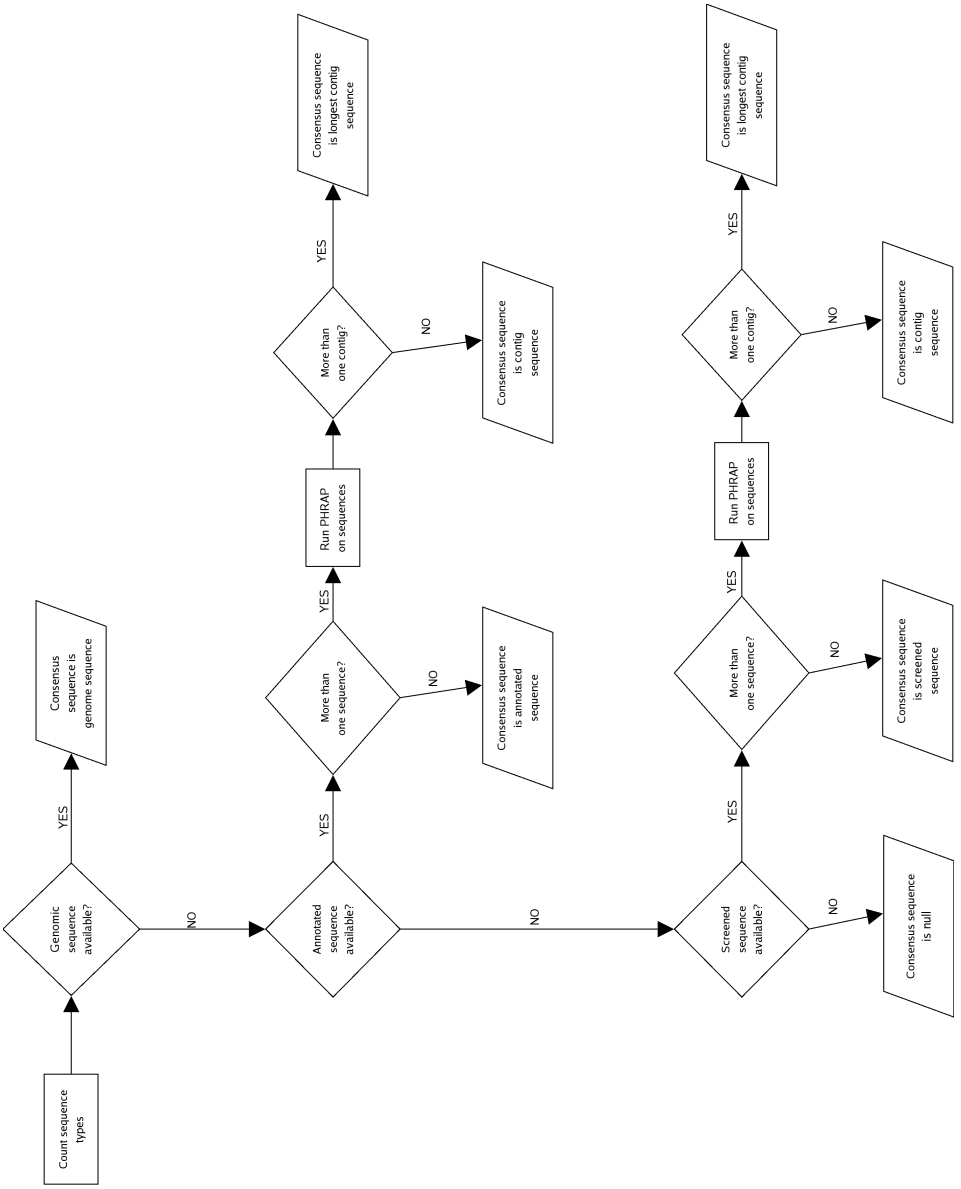


Figure 2.2: Rules for building consensus sequences

First, TaxMan looks for a sequence from a fully sequenced genome. If this is available, it is assumed to represent the full length of the gene and to be the highest-quality sequence available, and is used as the consensus sequence.

If there is no genome-derived sequence for the gene, TaxMan looks for regular annotated sequences. It is assumed that sequences which have been annotated with one of the known gene name synonyms are likely to be higher quality and more complete than screened sequences. If multiple annotated sequences are found for a gene in a particular species, all annotated sequences are written out to a FASTA format (web reference 8) file and phrap (Gordon, Abajian and Green 1998) is used to cluster them into a consensus sequence. Using phrap ensures that overlapping sequences are combined to form the longest possible consensus sequence, and that, where possible, positions are covered by multiple reads to increase quality. If phrap cannot align all input sequences to form a single output sequence, multiple contiguous sequences (contigs) may be produced. This will be the case if, for instance, the 3' and 5' portions of a gene are represented in the sequence set but the central portion is absent. In such cases the longest contig produced by phrap is chosen as the consensus sequence.

If there are no annotated sequence available for the gene, TaxMan will look for screened sequences – sequences identified on the basis of BLAST similarity to a known sequence. Screened sequences are regarded as less reliable than annotated sequences; most of them will be EST sequences, which are known to have a relatively high error rate and have a maximum read length of ~700 bases, making them partial sequences in many cases. As is the case for annotated sequences, multiple screened sequences are clustered using phrap and a consensus sequence derived in the same

way. If no sequence data are available for a particular gene in a species, the consensus sequence is flagged as absent. This rule-based approach generates the optimum consensus sequence for phylogenetic analysis, in terms of completeness and quality.

2.3.5 Alignment

To facilitate rapid production of concatenated multiple sequence alignments for subsets of genes and taxa, TaxMan stores aligned protein (where applicable) and nucleotide sequences for each consensus sequence. The nature of the datasets dealt with by TaxMan presents several distinct difficulties for multiple sequence alignment.

- Because TaxMan is designed for analysis of large taxonomic groups, sequences for a given gene may display a high degree of divergence due to the large evolutionary distance between them. Because of the much higher degree of conservation and larger alphabet of protein sequences, alignments made at the protein level are more reliable for distantly related sequences.
- For any given gene, the dataset of sequences (DNA or protein) to be aligned may contain a mixture of full-length and partial sequences. This can be a challenge to many alignment algorithms, which assume global similarity of sequences. Methods that carry out local alignment are better at dealing with partial sequences, allowing for missing data at the ends of shorter sequences.
- Multiple alignment algorithms scale in different ways. The large datasets used by TaxMan mean that there may be many sequences (e.g. $>10^3$) for a given

gene. This presents an obstacle of computational complexity, requiring alignment methods that scale well to cope with large input datasets.

For maximum accuracy, multiple sequence alignment of protein coding genes is carried out in TaxMan using a multi-step strategy. A local alignment algorithm, Partial Order Alignment (POA) (Lee, Grasso and Sharlow 2002) is used. First, full-length, annotated sequences (those that end with a stop codon) are translated to give full-length protein sequences, which are aligned using POA. Then, the aligned protein sequences are used as a scaffold to align the corresponding DNA sequences using tranalign from the EMBOSS package (Olson 2002). Finally, any screened sequences are profile-aligned using POA to the aligned DNA sequences. Aligned DNA and protein sequences are stored in the database. This approach exploits the increased alignment accuracy of protein sequences while allowing phylogenetic analysis to be carried out on DNA sequences. By storing both aligned protein and aligned DNA sequences, TaxMan can produce alignments in either alphabet for phylogenetic analysis. For RNA genes, POA is used to align all DNA sequences simultaneously, and only aligned DNA sequences are stored.

2.3.6 Slicing

An important concept of the TaxMan environment is the "slice". A slice is a subset of genes and taxa that defines a supermatrix of aligned sequences (Sanderson *et al.* 2003; Yan, Burleigh and Eulenstein 2005). The concatenated matrix of taxa and genes defined by a slice can be output by TaxMan in formats suitable for phylogenetic

analysis very rapidly. A core principle of TaxMan is that the same dataset can be used to address different phylogenetic questions, or the same question in multiple ways, by examining subsets of genes and taxa. TaxMan lets the user define a slice in several different ways:

- By providing lists of gene names and NCBI taxids
- By automatically building slices for each gene that include the gene and all taxa that have sequence data for that gene
- By selecting a defined number of representative species from each group at a given taxonomic level for a user-defined set of genes. Representative species are chosen in order of sequence completeness for the given list of genes, so that missing data in the final slice is minimised. For example, the user could specify a slice that included one representative species from each genus, or 5 representative species from each class, depending on the level of the phylogenetic relationship under investigation.

TaxMan produces concatenated multiple sequence files in NEXUS (Maddison, Swofford and Maddison 1997), FASTA (web reference 8) and PHYLIP (Felsenstein 2005) formats, suitable for analysis in a range of phylogenetic software. The NEXUS file format allows inclusion of character sets to define partitions. TaxMan takes advantage of this by including character sets to define individual genes and codon positions in outputted NEXUS files. This facilitates the use of complex, partitioned models in phylogenetic analysis, a key factor in accurate reconstruction (see Section 2.5.5 for a more detailed discussion).

2.3.7 Storing trees

TaxMan stores phylogenetic trees resulting from analysis of alignment slices. Trees in Newick (web reference 9) and NEXUS (Maddison, Swofford and Maddison 1997) format can be read. To facilitate retrieval of trees, individual nodes are stored as records in a database table. For each node, links are stored to the parent and child nodes (thus storing the structure of the tree) along with associated data like branch length and support values. Importantly, NCBI taxids are stored for the terminal nodes. This lets TaxMan retrieve information about the terminal taxa when producing trees for viewing, allowing, for instance, terminal nodes to be labelled with taxids, common names, the order to which each species belongs, etc.

TaxMan also parses the NCBI taxonomy dump file and stores the nodes in the same fashion. Because phylogenetic trees are stored as nodes, with terminal nodes annotated with taxids, TaxMan can extract the corresponding nodes from the NCBI taxonomy and output a 'pruned' taxonomy, showing just the relationships between the species included in the phylogenetic tree. This can assist in comparison of phylogenetic trees to taxonomic hypotheses. TaxMan associates trees with slices. For any given user tree stored in the TaxMan database, there will be an associated slice, making it easy to identify which genes the tree was built from.

2.4 Implementation

The implementation of TaxMan can be described in three parts; the code, the database and the external programs.

2.4.1 Perl + Bioperl + modules

TaxMan is written in Perl (web reference 10). Perl has a long and distinguished history as a language for bioinformatics, mainly due to its excellent text-processing abilities (particularly regular expressions) and shallow learning curve. Perl's flexible approach means that it is suitable for both small, one-off scripts and larger, more complex projects. Programs written in Perl can use 'modules' - libraries of code that carry out common tasks. This is important in large software projects for two reasons: first, it allows the programmer to make use of pre-existing code; second, it allows custom libraries to be written, encouraging code reuse and efficient programming. TaxMan takes advantage of pre-existing code by using modules to include functionality that would otherwise have to be written from scratch. The BioPerl project (Stajich *et al.* 2002) is a large collection of modules that carry out common bioinformatics tasks such as reading sequence files, parsing BLAST reports and manipulating trees. BioPerl is particularly valuable for its ability to assist in reading many common sequence and alignment file formats, a task that would otherwise require a great deal of repetitive coding. TaxMan also makes use of non-bioinformatics related Perl modules to carry out certain tasks. This include handling database connections and drawing the user menu interface.

2.4.2 Databasing

All types of data handled by TaxMan – sequences, species data, alignments and trees – are stored in a relational database. The relational database management system

(RDBMS) PostgreSQL (web reference 11) is used to manage the database. Whenever large volumes of data are required to be stored, a relational database can be useful. It allows the programmer to specify relationships between different types of data (for example, species, taken from the NCBI taxonomy, and sequences, taken from GenBank, both have a taxid field, so they are explicitly related). Additionally, the RDBMS takes care of indexing and searching the data, resulting in very fast storage and retrieval of records. This allows TaxMan to run quickly even when manipulating very large datasets.

2.4.3 External programs

TaxMan relies on a number of external bioinformatics tools to carry out various steps in the phylogenetics pipeline. These are the same tools that might be used by a researcher carrying out a large-scale phylogenetic project manually; in TaxMan, however, they are run 'behind the scenes' with no direct user input. The individual tools are listed below

- BLAST (Altschul *et al.* 1997) is a sequence similarity search tool that takes a query sequence and identifies any sequences in a database file to which the query sequence has significant similarity. When extracting screened sequences, BLAST is used to identify candidate sequences that share similarity with a database of known genes.
- Phrap (Gordon, Abajian and Green 1998) is a tool for assembling a collection of overlapping sequences into contiguous sequences. Given a set of input

sequences that display a certain level of similarity, phrap will produce a contiguous ('contig') output sequence that will be a consensus of the input sequences. Because the consensus sequence will usually be longer and less error-prone than the input sequences, phrap is used to build consensus sequences in TaxMan for genes in species where there are multiple annotated or EST sequences available.

- POA (Lee, Grasso and Sharlow 2002) is a multiple sequence alignment program that uses partial order graphs to align sets of sequences. It carries out local rather than global alignment and as such is capable of aligning sets of sequences of varying length. This property is important in the TaxMan pipeline, since the set of sequences to be aligned for a single gene may contain partial sequences (perhaps derived from ESTs) along with full-length sequences (perhaps derived from fully sequenced genomes).
- Tranalign is part of the EMBOSS package (Olson 2002). Given a set of aligned protein sequences and a set of corresponding DNA sequences, it produces a set of aligned DNA sequences using the protein sequences as a scaffold. Tranalign is a core component of the TaxMan alignment strategy, since alignment at the protein level is more accurate than alignment at the DNA level, especially for highly divergent sequences such as are likely to be found in a TaxMan dataset.

2.5 Features & Discussion

2.5.1 Searching in Genbank files

TaxMan Rapidly searches within GenBank files for DNA sequences of interest based on annotation. To accomplish this, it takes advantage of the GenBank file format which allows sophisticated annotation of nucleotide sequences. A single file may contain multiple independent GenBank records. Each GenBank record refers to a single contiguous DNA sequence and contains metadata – the source of the sequence, the locus name, any publications it is associated with, etc. Additionally, each record contains one or more features – regions within the sequence that are of biological interest. Figure 2.3 gives a graphical representation of the GenBank file format.

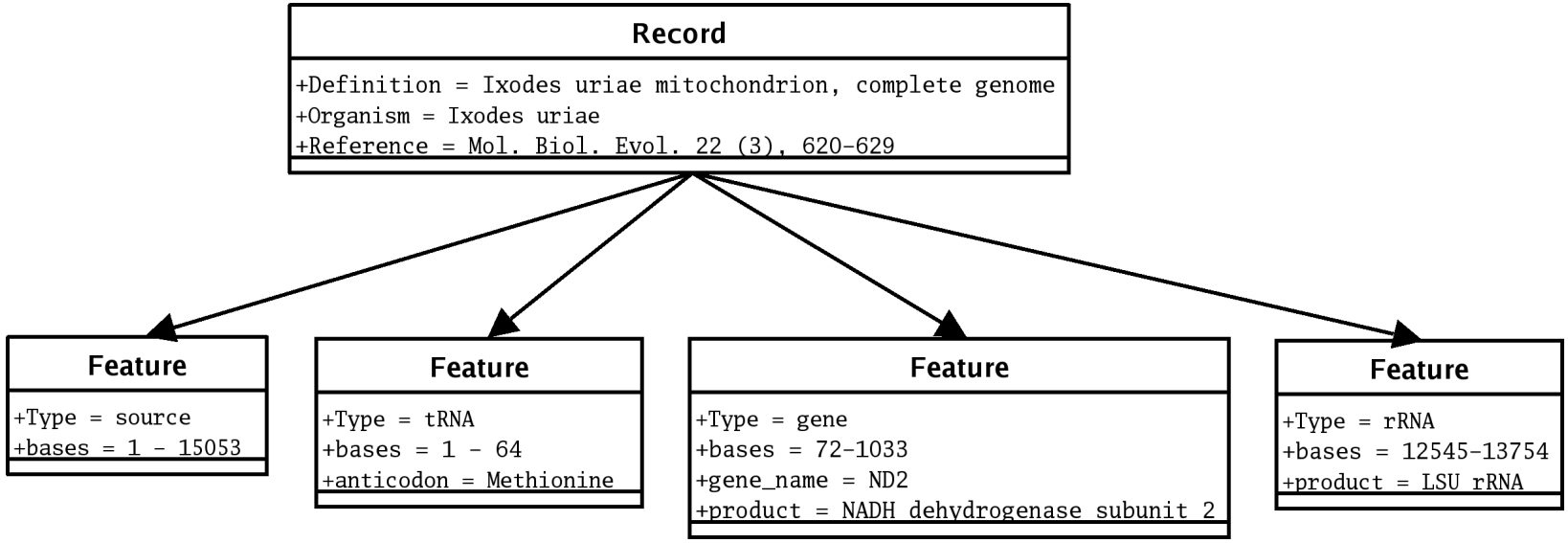


Figure 2.3: Graphical representation of a GenBank record

All records contain, minimally, a 'source' feature pertaining to the whole sequence. Other features can include genes, coding regions, promotors and transcribed regions (mRNAs). Records have a 'primary' tag which defines the type of feature and additional tags that carry feature metadata. The additional tags are different for each feature type (see Figure 2.3).

TaxMan uses the 'primary tag' to identify features that may be of interest, and uses the metadata tags to determine whether a feature should be extracted on the basis of annotation. Briefly, to be considered for inclusion in a dataset, a feature must satisfy the following requirements:

- Type is 'CDS' or 'rRNA'
- The feature has a 'gene' tag or a 'product' tag

If these conditions are satisfied, TaxMan compares the value of the 'gene' or 'product' tag to the predefined list of gene names and synonyms. If the tag matches one of the synonyms, the sequence of that feature is extracted and entered into the sequences table in the database with the appropriate gene name.

This approach to sequence acquisition allows very large input datasets to be processed rapidly. The criteria for consideration ensure that records without any features of interest (for example, those containing only tRNAs or unanotated genomic sequence) can be ignored, while ensuring that records containing multiple pertinent features (for example, annotated whole mitochondrial genomes) are exhaustively processed. The usefulness of this approach rests on two assumptions; first, that the majority of GenBank records are correctly annotated; second, that the user has assembled a comprehensive list of synonyms for each gene of interest. TaxMan assists with the

second assumption by parsing a collection of GenBank records and reporting any commonly-occurring gene names that are potential unrecognised synonyms. Currently, however, the approach is vulnerable to misannotated features that could be erroneously incorporated into a dataset.

2.5.2 Searching datasets for sequences of interest

As well as extracting sequences based on GenBank annotation, TaxMan also supports extraction of sequences based on BLAST similarity to known sequences, increasing the amount of data that can be collected. Two common scenarios for searching datasets are as follows:

- The user wants to use sequence data from expressed sequence tag (EST) projects. Since ESTs are single reads of inserts derived from random mature mRNAs, they often contain coding sequence that is useful in a phylogenetic context, but they are generally unannotated. Sequence similarity allows TaxMan to identify relevant sequences from this type of data.
- The user wants to incorporate sequence data from GenBank records that were unannotated, missanotated or were annotated with an unknown synonym, and thus would not be added on the basis of annotation.

TaxMan uses a database of known genes, formatted for use by BLAST, to identify useful sequences. Each input sequence is used the query in a BLASTX search against a database of known protein-coding genes and a BLASTN search against a database of known rRNA genes. A variable high E-value cutoff is used to include only sequences with highly significant sequence similarity to a known sequence. This strategy

assumes:

- All the sequences in the BLAST database are correctly annotated
- All sequences in the input dataset are correctly assigned to a species in the NCBI taxonomy
- The genes of interest are single-copy; any sequence with significant sequence similarity to one of the known genes is orthologous to it

Depending on the size of the input dataset to be searched, and of the BLAST databases, similarity searching can be the most time-consuming part of the TaxMan pipeline. However, BLAST searches of multiple sequences can easily be distributed across multiple computing nodes, or across multiple CPUs on a single computer, making this step a candidate for parallel processing.

2.5.3 Constructing consensus sequences

Species-level phylogenetic analysis requires that each species in an alignment is represented by a single sequence. However, the TaxMan approach to data gathering ensures that for some genes in some species, multiple sequences will be available. For example, a mitochondrial gene might be represented in a given species by (1) a sequences extracted from a whole mitochondrial genome sequence, (2) a collection of short annotated sequences collected for population genetic studies, and (3) a number of sequences from an EST project that have been extracted on the basis of BLAST similarity. TaxMan builds consensus sequences in such cases, using a rule-based approach that embodies *a priori* assumptions about the likely quality and completeness of each type of sequence.

The consensus-building step is the only stage in which sequences derived from fully-sequenced genomes are treated differently to other annotated sequences. Sequences derived from genomes (using annotated GenBank features) are assumed to be full length sequences, and to be higher quality than other types of sequences due to the multiple coverage of genome sequencing projects. They are also assumed to have the most accurate annotation. Annotated sequences that are not derived from whole genome sequences are less likely to be full length and high quality than those that are. Screened sequences, which have been added to the dataset on the basis of sequence similarity, are assumed to be likely to be partial and error prone, since many screened sequences are ESTs. Additionally, the gene name is assumed to be less reliable than that of annotated sequences since it is allocated indirectly.

2.5.4 Multiple sequence alignments and storing aligned sequences

The multiple sequence alignment strategy implemented in TaxMan is designed to ensure that both aligned DNA and aligned protein sequences are available, where possible, for analysis while maintaining a high level of alignment accuracy. To this end, multiple sequence alignment is carried out at the protein level for those sequences for which a complete protein translation is possible. The protein alignment is then used as a template to align the corresponding DNA sequences. Sequences that cannot be translated fully (screened sequences, for example, ESTs) are profile-aligned to the annotated DNA sequences. (see figure 2.4).

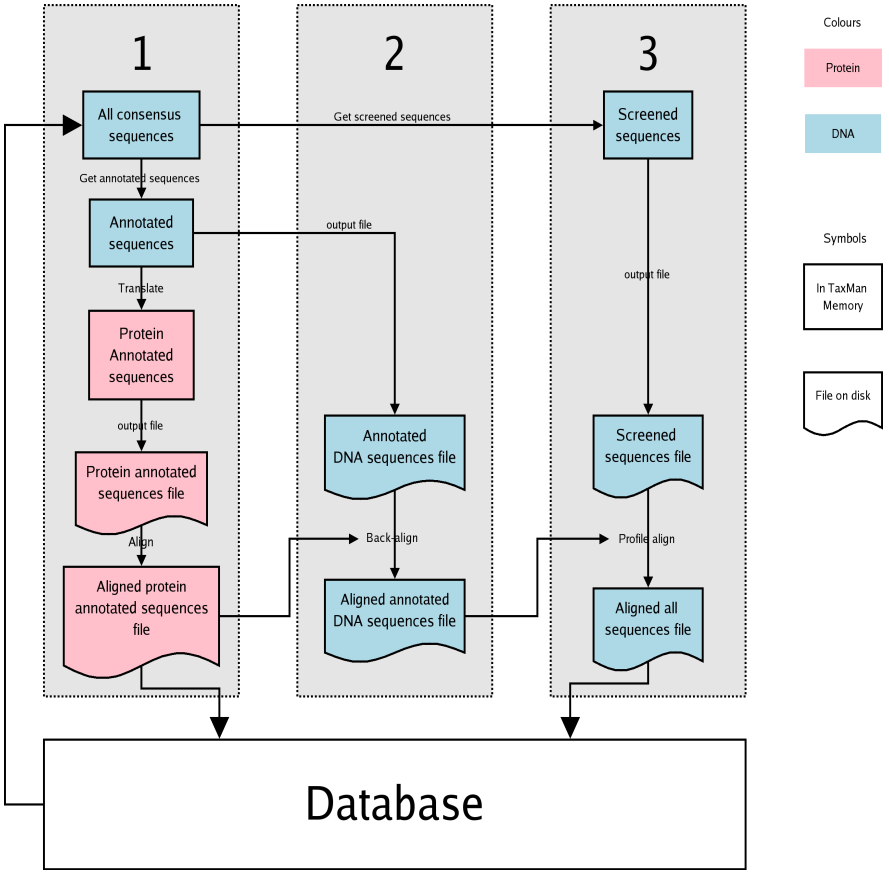


Figure 2.4: Flowchart showing the alignment process used in TaxMan

1 – Consensus DNA sequences built from annotated sequences are extracted from the database, translated into protein sequences and the protein sequences aligned

2 – The consensus DNA sequences extracted in step 1 are aligned using the protein alignment as a guide

3 – The consensus DNA sequences built from screened sequences are extracted from the database and profile-aligned to the DNA sequences from step 2

Aligned protein and DNA sequences are stored in the database

Alignment is more accurate at the protein level than the DNA level. This is due to the larger alphabet of amino acids compared to nucleotides. In particular, for highly divergent sequences, similarity is likely to be preserved at the amino acid level when none is detectable at the nucleotide level.

A particular problem to be overcome when assembling large sequence datasets from public databases is that of partial sequences. For any given gene, consensus sequences built from EST sequences are likely to be shorter than consensus sequences built from genome sequences. In such a scenario, a multiple sequence alignment algorithm must be used that can carry out local alignment for the partial sequences, allowing them to align with the portion of the full-length sequences to which they are homologous. In particular, any algorithm that penalises end-gaps will perform poorly on partial sequences. The alignment algorithm used in TaxMan, POA, uses partial graphs to perform local multiple sequence alignment, giving good results on collections of sequences of varying length.

2.5.5 Producing multigene partitioned alignments

An important feature of TaxMan is the ability to produce alignments containing multiple genes for a given set of taxa, with character sets and partitioning schemes corresponding to meaningful biological categories. When TaxMan produces an alignment in NEXUS file format, it includes the following character sets:

- For each gene, a set containing every character in the gene
- For each protein-coding gene gene, three sets each containing every character

in the gene that is in a certain codon position

- For each codon position, a set containing all characters in that codon position across all protein-coding genes

These predefined character sets make it easy for the user to manipulate the alignment in NEXUS-compatible software packages (e.g. PAUP* (Swofford 2006) and MrBayes (Ronquist and Huelsenbeck 2003)). For example, excluding all third codon position sites from the ND1 gene while including all others in an analysis is straightforward:

```
exclude ND1.codon_position_3;
```

as is including only RNA genes from an alignment of all mitochondrial genes:

```
exclude all; include RNA_12S RNA_16S;
```

TaxMan also automatically defines two partitioning schemes, by_gene (in which each individual gene has its own partition) and by_codon (in which each codon position has its own partition). Other partitioning schemes are easy for the user to set up, using the named character sets already present in the file. For example, to allow each gene to have an independent gamma rate variation parameter:

```
set partition = by_gene;
unlink shape=(all);
```

The use of a partitioned model has been found to be of crucial importance in accurate phylogenetic reconstruction using multiple concatenated genes (see Chapter 3). Partitioned evolutionary models allow different models and/or model parameters to be applied to different sets of characters (for example, different genes) and has been shown to be an important component of an accurate model.

2.5.6 Storing and retrieving trees resulting from phylogenetic analysis

Most phylogenetics programs produce output trees in newick or NEXUS format. In these two similar formats the tree is represented as a text string in which pairs of parentheses define the branch structure. TaxMan reads trees in newick and NEXUS format, but stores them internally as collections of nodes. Each node in a tree has 'parent' and 'children' fields, which define the structure of the tree, along with metadata such as branch lengths and support values. Importantly, storing nodes in this manner means that large taxonomic trees (such as the NCBI taxonomy) can be stored using the same framework. This means that phylogenetic and taxonomic trees can be manipulated in the same way. For example, when TaxMan produces an annotated copy of a stored user tree (for viewing in a tree viewing application) it can also produce a pruned copy of the NCBI taxonomic tree, showing only the species included in the stored user tree, for comparison. Because the pruned taxonomic tree is in the same format as the stored user tree and contains the same terminal taxa, the two can be easily compared in standard tree-viewing software.

Storing trees is useful for the user since TaxMan also stores metadata about the tree-building event. When storing a tree produced using MrBayes, TaxMan stores (1) the MrBayes command string that was used to carry out the analysis and (2) the parameter string describing estimates of the evolutionary parameters used in the analysis. Additionally, TaxMan associates the tree with the slice from which it was built, ensuring that data provenance is maintained. For a given tree, the user can use TaxMan to trace the flow of data back to the slice, to the aligned sequences, the

consensus sequences and finally back to the accession numbers of the GenBank records from which sequence data was originally extracted.

2.6 Benchmarking

TaxMan was used to assemble two large sequence datasets. A dataset comprising 15 mitochondrial genes along with the nuclear large and small subunit RNA genes was assembled for the Chelicerata (details of analysis in Chapter 3). A dataset comprising 15 mitochondrial genes, nuclear large and small subunit RNA genes and four nuclear protein-coding genes was assembled for the Lophotrochozoa.

2.6.1 Chelicerate dataset

To assemble the chelicerate dataset, NCBI Entrez was used to download all NCBI nucleotide records for the class Chelicerata in GenBank format. In total, ~82,000 records were retrieved. ~12,000 were annotated sequences from the CoreNucleotide database and ~70,000 were EST sequences from the EST database. GenBank was also searched to identify the following gene name synonyms for the genes of interest (Table 2.2).

Canonical name	Synonyms
ATP6	ATPase6,ATP6,ATPase 6,atp6
ATP8	atp8,ATPase8,ATP8,ATPase 8
COX1	COI,coxi,coxI,COX1,CO1,COXI,co1,coi,cox1, cytochrome oxidase subunit 1,coI,cytochrome oxidase subunit I
COX2	co2,COXII,cox2,coii,COII,COX2,coxii,CO2, cytochrome oxidase subunit 2, cytochrome oxidase subunit II
COX3	coiii,COX3,COIII,COXIII,CO3,coxiii,co3,cox3, cytochrome oxidase subunit 3
CYTB	cytb,cob,CYTB,COB,cytochrome b, cytochrome b protein
ND1	ND1,nad1,NAD1,nd1,NADH dehydrogenase subunit 1,NADH dehydrogenase 1
ND2	nad2,nd2,NAD2,ND2,NADH dehydrogenase subunit 2,NADH dehydrogenase 2
ND3	nad3,ND3,NAD3,nd3,NADH dehydrogenase subunit 3,NADH dehydrogenase 3
ND4	NAD4,ND4,nd4,nad4,NADH dehydrogenase subunit 4,NADH dehydrogenase 4
ND4L	NAD4l,ND4L,nad4L,nad4l,nd4L,nd4l,NAD4L, ND4l,NADH dehydrogenase subunit 4L, NADH dehydrogenase 4L
ND5	nd5,NAD5,ND5,nad5,NADH dehydrogenase subunit 5,NADH dehydrogenase 5
ND6	nd6,nad6,NAD6,ND6,NADH dehydrogenase subunit 6,NADH dehydrogenase 6
RNA_12S	12s rRNA,12S rRNA,12S ribosomal RNA, s-rRNA
RNA_16S	16S rRNA,16s rRNA,16S ribosomal RNA, l-rRNA
RNA_LSU	28S ribosomal RNA,28S large subunit ribosomal RNA,28S rRNA
RNA_SSU	18S ribosomal RNA,18S small subunit ribosomal RNA,18S rRNA

Table 2.2: Gene names and synonyms used to gather the Chelicerata dataset

In total, TaxMan extracted 16,103 sequences of interest from the input set of GenBank records. 7,294 sequences were extracted on the basis of annotation, 463 of which were extracted from records representing whole mitochondrial genome sequences and 6,831 of which were extracted on the the basis of annotation from non-genome records. Using a BLAST database containing only full-length sequences from species which had a fully sequenced mitochondrial genome, 8,809 sequences were extracted on the basis of sequence similarity. The distribution of sequence data was uneven both across species (and higher level taxonomic groups) and across genes. The number of sequences per gene ranged from 5,613 for RNA_16S to 45 for ATP8. The number of sequences per species ranged from 1,697 for *Amblyomma americanum* (a species with a large EST sequencing project) to only a single sequence for 358 species. In total, 1,506 species from 19 orders were represented.

In the consensus-building stage, 3,510 consensus sequences were built (see Table 2.3). 270 were built from genome-derived sequences, 2,917 from annotated sequence and 323 from screened sequences (mostly ESTs). The relatively small contribution made by screened sequences results from the fact that EST projects are most likely to be carried out on well-studied species, which are also most likely to have complete mitochondrial genome sequences or large numbers of annotated sequences. Since genome and annotated sequences take precedence over screened sequences in the consensus building process, large numbers of screened sequences will not be used. For example, 957 screened sequences were found for the 16S ribosomal RNA gene in the tick *Amblyomma americanum*, but were not used as an annotated sequence was also available.

In common with the extracted sequences, distribution of data across consensus sequences was uneven. The number of consensus sequences per gene ranged from 788 (RNA_16S) to 22 (ND6) and the number of consensus sequences per order ranged from 1,782 (for Araneae) to 2 (for Palpigradi). The process of assembling the dataset, including sequence gathering, consensus building and alignment, took approximately 6 hours using a desktop computer with a 2.8 Ghz processor and 2 Gb RAM.

Order	no. taxa	RNA_16S	COX1	RNA_LSU	RNA_SSU	ND1	RNA_12S	EF1A	ND3	COX2	ATP6	ND5	COX3	CYTB	ND2	ND4	ATP8	ND4L	ND6	*All genes*
Araneae	710	472	433	266	73	302	114	69	4	6	7	5	5	5	5	6	4	3	3	*1782*
Ixodida	208	141	52	62	57	45	160	0	29	15	15	21	16	20	15	15	11	15	14	*703*
Scorpiones	122	116	50	20	16	2	17	0	2	5	3	2	3	2	3	2	3	2	2	*250*
Mesostigmata	100	12	26	54	55	1	8	22	1	1	1	1	1	1	1	1	1	1	1	*189*
Trombidiformes	97	2	63	14	40	1	9	1	1	1	1	1	1	1	1	1	1	1	1	*141*
Oribatida	97	0	13	81	16	0	0	1	0	0	0	0	0	0	0	0	0	0	0	*111*
Astigmata	60	32	16	18	11	0	5	0	0	2	2	2	1	1	1	1	0	0	0	*92*
Opiliones	39	1	4	34	38	0	0	2	0	0	0	0	0	0	0	0	0	0	0	*79*
Pantopoda	37	5	7	23	30	0	0	0	0	1	1	0	1	0	1	0	1	0	0	*70*
Xiphosura	4	4	4	2	2	1	1	1	1	1	1	1	1	1	1	1	1	1	1	*26*
Uropygi	4	1	1	2	4	0	0	1	0	1	1	0	1	0	1	0	1	0	0	*14*
Holothyrida	5	1	0	4	5	0	1	0	0	0	0	0	0	0	0	0	0	0	0	*11*
Pseudoscorpiones	9	0	5	2	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	*10*
Amblypygi	3	0	1	1	2	0	0	0	0	1	1	0	1	0	1	0	1	0	0	*9*
Opilioacarida	2	1	1	1	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	*6*
Solifugae	3	0	0	3	3	0	0	0	0	0	0	0	0	0	0	0	0	0	0	*6*
Endeostigmata	3	0	0	1	2	0	0	1	0	0	0	0	0	0	0	0	0	0	0	*4*
Ricinulei	2	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	*3*
Palpigradi	1	0	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	*2*
All orders	1506	788	676	590	362	353	315	98	38	34	33	33	31	31	30	27	24	23	22	*3508*

Table 2.3: Numbers of consensus sequences by gene for each chelicerate order

TaxMan was used to automatically specify a slice consisting of all 17 genes and one representative species from each order. The resulting alignment was 22,188 characters long and each species had, on average, 7787 characters (~35%) present.

2.6.2 Lophotrochozoa dataset

To assemble the lophotrochozoan dataset, a different approach was taken. The entire GenBank archive was downloaded from the NCBI FTP site, along with the taxonomy dump. Since 'Lophotrocozoa' does not exist as a node in the NCBI taxonomy, a custom Perl script was then used to identify a list of species-level taxon IDs belonging to lophotrochozoan phyla and extract only records belonging to those species. 530,166 GenBank records were retrieved. Records pertaining to environmental samples or unclassified species were discarded. When obtaining records using Entrez, records from each section of GenBank can be downloaded separately. This is helpful, as it allows records from the CoreNucleotide section (mostly representing annotated sequence) and records from the EST section (representing unannotated sequences) to be processed by TaxMan independently. For example, one can choose not to look for annotated sequences in the EST record set, since none are expected to be found. Because the Lophotrochozoa record set was not obtained through Entrez, this option was not available, therefore all records were processed together. The set of gene name synonyms used is shown in Table 2.4.

Canonical name	Synonyms
ATP6	ATPase6,ATP6,ATPase 6,atp6
ATP8	atp8,ATPase8,ATP8,ATPase 8
COX1	COI,coxi,coxI,COX1,CO1,COXI,co1,coi,cox1,cytochrome oxidase subunit 1,coI,cytochrome oxidase subunit I
COX2	co2,COXII,cox2,coii,COII,COX2,coxii,CO2,cytochrome oxidase subunit 2,cytochrome oxidase subunit II
COX3	coiii,COX3,COIII,COXIII,CO3,coxiii,co3,cox3,cytochrome oxidase subunit 3
CYTB	cytb,cob,CYTB,COB,cytochrome b,cytochrome b protein
ND1	ND1,nad1,NAD1,nd1,NADH dehydrogenase subunit 1,NADH dehydrogenase 1
ND2	nad2,nd2,NAD2,ND2,NADH dehydrogenase subunit 2,NADH dehydrogenase 2
ND3	nad3,ND3,NAD3,nd3,NADH dehydrogenase subunit 3,NADH dehydrogenase 3
ND4	NAD4,ND4,nd4,nad4,NADH dehydrogenase subunit 4,NADH dehydrogenase 4
ND4L	NAD4l,ND4L,nad4L,nad4l,nd4L,nd4l,NAD4L,ND4l,NADH dehydrogenase subunit 4L NADH dehydrogenase 4L
ND5	nd5,NAD5,ND5,nad5,NADH dehydrogenase subunit 5,NADH dehydrogenase 5
ND6	nd6,nad6,NAD6,ND6,NADH dehydrogenase subunit 6,NADH dehydrogenase 6
RNA_12S	12s rRNA,12S rRNA,12S ribosomal RNA,s-rRNA
RNA_16S	16S rRNA,16s rRNA,16S ribosomal RNA,l-rRNA
RNA_LSU	28S ribosomal RNA,28S large subunit ribosomal RNA,28S rRNA
RNA_SSU	18S ribosomal RNA,18S small subunit ribosomal RNA,18S rRNA
EF1A	EF-1alpha,EF-1a,EF1a,EF1-alpha,elongation factor-1 alpha,elongation factor 1 alpha
H3	histone H3,H3
ACTIN	actin

Table 2.4: Gene names and synonyms used to gather data for the Lophotrochozoa dataset

In total, TaxMan extracted 76,790 lophotrochozoan sequences from the input set of GenBank records. 31,454 sequences were extracted on the basis of annotation, 731 of which were from records containing fully sequenced annotated mitochondrial genomes. The majority of annotated sequences (30,723 [97.7%]) were extracted from non-genome records. Using a BLAST database containg only full-length sequences from species which had a fully sequenced mitochondrial genome (but including nuclear genes for those species), 45,336 sequences were extracted on the basis of sequence similarity. The uneven distribution of sequence data, which is likely to be characteristic of large taxonomic groups, was again seen. The number of sequences per gene ranged from 20,728 (RNA_LSU) to 95 (ATP8). The number of sequences per species ranged from 8,644 for *Schistosoma mansoni* (a human trematode parasite with a mature genome project) to a group of 1,524 species for which there was only a single gene. In total, 8097 species had sequence data, representing 135 orders and 30 classes.

In the consensus-building stage, 14,387 consensus sequences were built. The majority (12,819) were derived from annotated non-genome sequences. Of the remainder, 561 were built from annotated genome sequences, and 1,007 were built from screened sequences extracted on the basis of sequence similarity. As in the chelicerate dataset, screened sequences made a much smaller contribution to the set of consensus sequences than might have been expected from raw numbers. The number of consensus sequences per gene ranged from 3,512 (COX1) to 42 (ATP8). The number of consensus sequences per class ranged from 6,473 for Gastropoda to 3 for Stenolaemata (see table Table 2.5).

The whole process took approximately 72 hours on a desktop computer with a 2.8 Ghz processor and 2 Gb RAM. The extremely large numbers of sequences and large degree of divergence meant that the usual TaxMan alignment strategy was unsatisfactory. Alignment of all sequences for each gene was extremely time consuming and the resulting alignments were poor. Aligning just the sequences in a slice, following definition of a slice, gave improved alignments; this approach was used for all analyses of the Lophotrochozoa dataset. It seems likely that the align-once strategy, while optimal for smaller datasets (<1000 sequences per gene), will be ineffective at dealing with large datasets, where a traditional alignment strategy must be used.

TaxMan was used to automatically specify a slice consisting of all 21 genes and one representative species from each class. The resulting alignment was 31,523 characters long and each species had, on average, 9456 characters (~30%) present.

2.6.3 Beta testing on additional datasets

TaxMan was tested by Chris Jiggins (University of Edinburgh) on a lepidopteran dataset and by Mark Blaxter (University of Edinburgh) on a coleopteran dataset. Jamie Floyd (University of Edinburgh MSc student) used TaxMan to assemble a dataset of arthropod mitochondrial genes to investigate hexapod phylogeny. Charlie Goodway (University of Edinburgh BSc student) used an early version of TaxMan to assemble an ecdysozoan dataset to investigate the phylogenetic position of Tardigrada. Feedback was encouraged and used to develop and introduce additional features. Joseph Hughes (University of Glasgow) used TaxMan to assemble a dataset of lice

sequences. TaxMan has also been available for download on the Blaxter Lab website (`www.nematodes.org`).

Containing group	class	no. taxa	COX1	RNA_16S	RNA_LSU	RNA_SSU	RNA_12S	ND1	H3	CYTB	ACTIN	EF1A	COX2	COX3	ND4L	ND4	ATP6	ND3	ND6	ND2	ND5	ATP8	*All genes*
Mollusca		4691	2797	2521	1366	934	530	330	215	208	158	56	75	88	72	62	35	31	42	33	39	19	*9611*
	Scaphopoda	28	15	2	6	19	0	2	2	2	0	0	2	2	2	2	2	2	2	2	2	2	*68*
	Aplacophora	9	5	5	8	6	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	*30*
	Cephalopoda	312	212	169	97	78	121	8	51	39	44	30	20	43	7	8	8	8	8	7	8	8	*974*
	Polyplacophora	35	30	27	29	28	0	0	28	0	1	1	0	0	0	0	0	0	0	0	0	0	*144*
	Gastropoda	3268	2017	1879	909	544	376	206	99	144	103	11	28	14	53	10	13	8	23	10	18	8	*6473*
	Bivalvia	1039	518	439	317	259	33	114	29	23	10	14	25	29	10	42	12	13	9	14	11	1	*1922*
Platyhelminthes		1392	229	50	870	817	35	61	15	22	4	42	18	19	13	20	24	26	14	21	16	0	*2316*
	Monogenea	277	20	11	161	142	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	*335*
	Turbellaria	252	35	7	88	196	1	4	14	7	0	12	6	7	0	2	5	0	1	3	4	0	*392*
	Trematoda	527	103	18	384	266	10	46	1	7	3	6	6	6	7	8	6	15	6	8	6	0	*912*
	Cestoda	336	71	14	237	213	24	11	0	8	1	23	6	6	6	10	13	11	7	10	6	0	*677*
Annelida		850	374	279	285	524	125	77	40	21	0	29	21	6	5	6	6	6	6	7	5	17	*1839*
	Clitellata	264	101	139	90	101	48	16	1	2	0	6	16	2	2	3	2	2	2	3	2	13	*551*
	Polychaeta	417	121	137	131	302	2	3	39	17	0	18	3	3	3	3	3	3	3	3	3	3	*800*
	Branchiobdellae	21	20	0	2	20	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	*42*
	Hirudinida	148	132	3	62	101	75	58	0	2	0	5	2	1	0	0	1	1	1	1	0	1	*446*
Bryozoa		93	22	68	10	22	10	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	*133*
	Phylactolaemata	12	0	10	0	7	9	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	*27*
	Stenolaemata	2	0	0	1	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	*3*
	Gymnolaemata	79	22	58	9	13	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	*103*
Nemertea		77	21	33	19	47	0	0	9	0	0	4	0	0	0	0	0	0	0	0	0	0	*133*
	Anopla	29	3	22	8	12	0	0	3	0	0	3	0	0	0	0	0	0	0	0	0	0	*51*
	Enopla	48	18	11	11	35	0	0	6	0	0	1	0	0	0	0	0	0	0	0	0	0	*82*
Brachiopoda		65	35	6	9	28	14	4	2	4	0	1	4	4	3	4	4	4	4	4	4	2	*140*
	Craniata	2	1	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	*2*
	Lingulata	6	3	0	1	2	0	1	0	1	0	1	1	1	1	1	1	1	1	1	1	1	*19*
	Phoroniformea	8	2	0	4	7	0	1	2	1	0	0	1	1	1	1	1	1	1	1	1	0	*26*
	Rhynchonellata	49	29	6	3	19	14	2	0	2	0	0	2	2	1	2	2	2	2	2	2	1	*93*
Sipuncula		34	14	1	27	29	0	0	25	1	0	1	1	1	0	0	0	0	1	0	0	1	*102*
	Phascolosomatidea	17	5	0	16	16	0	0	15	0	0	0	0	0	0	0	0	0	0	0	0	0	*52*
	Sipunculidea	17	9	1	11	13	0	0	10	1	0	1	1	1	0	0	0	0	1	0	0	1	*50*
Pogonophora		29	18	13	6	18	0	2	1	2	0	3	2	2	0	1	1	2	2	2	0	2	*77*
	Perviata	13	4	8	2	11	0	1	0	1	0	0	1	1	0	0	0	1	1	1	0	1	*33*
	Vestimentifera	16	14	5	4	7	0	1	1	1	0	3	1	1	0	1	1	1	1	1	0	1	*44*
Rotifera		53	39	20	15	25	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	*100*
	Bdelloidea	16	13	1	2	7	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	*23*
	Monogononta	35	26	19	13	16	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	*75*
	Seisonidea	2	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	*2*
None		11	2	1	5	8	0	1	2	1	0	2	1	1	1	1	1	1	1	1	1	1	*32*
	Echiura	7	1	1	3	4	0	1	1	1	0	2	1	1	1	1	1	1	1	1	1	1	*24*
	Entoprocta	4	1	0	2	4	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	*8*
	All classes	*7295*	*3551*	*2992*	*2612*	*2452*	*714*	*475*	*309*	*260*	*163*	*138*	*122*	*121*	*94*	*94*	*71*	*70*	*70*	*68*	*65*	*42*	*14483*

Table 2.5: Numbers of consensus sequences per class and higher group for the Lophotrochozoa dataset

Table 5 gives a summary of the benchmarking results

Dataset	*GenBank records*	*Sequences extracted*	*Consensus sequences*	*Species represented*
Chelicerata	82000	16103	3510	1506
Lophotorochozoa	530116	96970	14529	7353

Table 2.6: TaxMan benchmark summary

2.7 Discussion

TaxMan has been shown to be a powerful tool for rapidly assembling large datasets of aligned sequences for phylogenetic analysis. In the examples described above, it handled very large input datasets comprising several hundred thousand GenBank records and extracted many sequences containing thousands of phylogenetic characters. By searching an input dataset for common feature names prior to extracting sequences, TaxMan allows the user to build as complete a list as possible of gene name synonyms, ensuring that the majority of annotated sequences are correctly identified. By also extracting sequences on the basis of similarity, TaxMan tries to ensure that all relevant sequence data are collected. By automatically selecting species to be included in a slice TaxMan can generate alignments to answer phylogenetic questions at multiple levels. In the above example, analysis of the slices mentioned would generate phylogenetic data for the ordinal and class-level phylogeny of the Chelicerata and Lophotrochozoa respectively. The slice paradigm also allows the same dataset to be used to address multiple different phylogenetic problems, or to

compare the answers given by different subsets of the data (for example, protein-coding vs ribosomal RNA genes; nuclear vs mitochondrial genes).

2.7.1 Low but important contribution of screened sequences

In both of the datasets examined (see Section 2.6) EST sequences comprised the majority of GenBank records used as input, and a large proportion of the sequences extracted from GenBank files were screened sequences. However, these screened sequences made only a very small contribution to the final set of aligned consensus sequences. This is largely a feature of the chosen gene set – in both benchmark cases, the genes chosen for study were well-known and had been heavily used in phylogenetics in the past. As such, they are likely to have been sequenced in many taxa; thus species with EST sequencing projects are likely to also have annotated sequences available for these genes. The annotated sequences will be used during consensus-building in preference to the screened sequences, leading to the strikingly low contribution of screened sequences to consensus sequences.

This pattern is likely to be the case for any set of well-studied genes – if an organism is well-studied enough to have an EST sequencing project, it probably has annotated sequences available for a number of commonly-used genes, particularly for mitochondrial genes since many metazoan mitochondrial genomes have been sequenced. The only scenario in which this pattern would not be seen is if the input gene set were selected on a criterion other than common use in phylogenetics.

2.7.2 Alignment strategies for large datasets

In general, TaxMan follows the 'align once' strategy as outlined in the Introduction (Section 2.3.1). This allows aligned protein and DNA sequences to be stored for all sequences, and allows rapid production of alignment files for different slices. For some genes in some datasets (e.g. COX1 in the Lophotrochozoa dataset) the large number of sequences make this approach unfeasible. For most alignment algorithms, computing time and/or memory requirements increase polynomially with number of sequences. For example, for any algorithm requiring pairwise comparison of sequences $\mathbf{½n^2}$ comparisons must be made in order to align n sequences; thus, the algorithm has order $\mathbf{n^2}$. As a result, very large sets of sequences cannot be aligned in a reasonable time or with the memory limitations of the computer system. For the 3512 COX1 consensus sequences in the Lophotrochozoa dataset, POA (Lee, Grasso and Sharlow 2002), MUSCLE (Edgar 2004) and CLUSTALW (Thompson, Higgins and Gibson 1994) were all unable to align the full set of sequences on a compute server with 4 Gigabytes of RAM. This limitation is likely to be overcome in the future with advances in the field of multiple sequence alignment, and is likely to occur only in the very largest datasets. To work around the difficulty for the Lophotrochozoan analysis, subsets of sequences were aligned using POA for the four genes with the most sequence data (COX1, RNA_16S, RNA_SSU and RNA_LSU). An alternative strategy for ribosomal RNA genes would be to use a core database (e.g. Ribosomal Database Project II; Cole *et al.* 2006) for alignment sourcing.

2.7.3 Consequences of consensus-building strategy

One drawback of TaxMan's hierarchical approach to building consensus sequences is the possibility of discarding useful data under certain circumstances. This is a consequence of the preference for annotated over screened sequences, without regard to sequence length. Consider the scenario where, for a given gene in a given species, there is a partial annotated sequence available, along with a collection of screened sequences from an EST project that, if combined using phrap, would cover the whole length of the gene. Despite the potentially larger amount of sequence data contained in the screened sequences, they will be discarded in favour of the partial annotated sequence. An extension to the TaxMan consensus system could address this issue in one of several ways.

One solution would be to simply use all sequences (annotated and screened) to build a consensus sequence using phrap, possibly assigning a higher quality score to annotated sequences. While straightforward, this solution would greatly increase the processing time required to build consensus sequences, since phrap would have to be run in a much greater proportion of cases. For example, in the Lophotrochozoa dataset, phrap was used to construct a consensus sequence in 1,737 cases. Under the above scheme, the number would be 12,593. For the majority of consensus sequences (all those where the scenario doesn't apply) this would result in no improvement in the length or quality of consensus sequence. This would be the case even if the input sequence set were restricted to the longest annotated sequence, plus all screened sequences.

A second possible solution is to specify a length for each of the genes of interest. This could be entered in the configuration file along with the synonyms. In cases where

both annotated and screened sequences were available, phrap consensus-building would only be carried out if the longest annotated sequence was less than 80% (or some arbitrary fraction) of the full gene length. Testing and implementing these ideas is a future goal for TaxMan.

2.7.4 On the inclusion of local datasets

One of the guiding principles of TaxMan is that as much sequence data as possible can be included in a dataset. To that end, it allows the user to add EST sequences on the basis of BLAST annotation. However, there are several EST software packages capable of clustering and organising EST sequencing projects (Parkinson *et al.* 2004a; Pertea *et al.* 2003). Because these packages are dedicated to EST analysis they employ more sophisticated techniques than TaxMan, and are therefore able to supply higher-quality gene objects. In its current state, TaxMan does not allow direct import of genes from any of these EST analysis packages. In order to include such data, the user would have to supply TaxMan with the raw EST sequences (thereby foregoing the improvements offered by dedicated software) or convert the gene objects to GenBank format files. Collections of EST sequences are routinely processed and annotated using bioinformatics tools that are highly automated. One of the intentions in building TaxMan is that phylogenetic analysis will join the list of routine procedures that are carried out on EST libraries. An important future direction for TaxMan, then, is the addition of routines for integration with the databases produced by EST analysis software and inclusion of gene objects.

2.7.5 Limitations & extensions

In its current form, TaxMan is a useful tool, but it is limited, by design, in several ways.

Multiple sequences

TaxMan assumes that the genes of interest are single copy and that the intention is to assemble orthologous sequences for species-level phylogeny. These assumptions inform the consensus building strategy, which dictates that a single consensus sequence must represent a gene for each species. For this reason, TaxMan is unsuitable for carrying out phylogenetic investigations of gene families. One possible extension of TaxMan would be the ability to hold multiple sequences for a gene in each species, allowing paralogues to be included in the dataset and facilitating gene family phylogenetics.

By the same token, the requirement for a single consensus sequence for each gene in each species means that TaxMan is not suitable for barcode-type datasets, where the same locus is sequenced in multiple specimens of the same species. Under the current TaxMan scheme, a collection of barcode sequences would be collated into a consensus sequence, losing any information regarding differences between individuals or populations. Extending the functionality of TaxMan to allow multiple sequences per gene for a single species would remove this limitation.

Alignment issues

When dealing with large taxonomic groups, there can be a high degree of divergence between sequences for a given gene. As a result, multiple sequence alignment is

challenging, and many alignments have poorly-aligned regions. Programs such as Gblocks (Castresana 2000) can identify well-aligned regions (or 'blocks' of sequences) and remove poorly-aligned regions. Alignment quality can also be evaluated using an objective function (Thompson *et al.* 2001). Incorporation of such a process as an optional step in TaxMan would be a useful addition and might improve phylogenetic accuracy. Another useful feature would be the ability for a user to edit alignments or use custom alignment programs and parameters.

While the alignment strategy in TaxMan is intended to make use of protein translations wherever possible, there are large numbers of sequences (for example, those taken from EST projects) for which this is not possible. Software packages for accurately translating EST sequences are available (ESTScan [Iseli, Jongeneel and Bucher 1999], DECODER [Fukunishi and Hayashizaki 2001]) and have been incorporated into a sophisticated EST translation pipeline (prot4EST [Wasmuth and Blaxter 2004]). Incorporation of the prot4EST pipeline into the TaxMan alignment process would allow a greater proportion of sequences to be aligned at the protein level, increasing alignment accuracy.

When extracting sequences on the basis of annotation, TaxMan relies on the feature names supplied by the GenBank record submitter, and on the synonyms supplied by the user. As a consequence, it is vulnerable to misannotated features or incorrect synonyms. It should be possible to flag most incorrectly identified sequences by viewing the alignment; however, such sequences may obscure genuine phylogenetic information contained in correctly-identified sequences.

Parallel processing

Many of the steps in the TaxMan pipeline are computationally intensive for large datasets and thus are good candidates for parallel processing. Making it possible for the TaxMan code to run on multiple processors would speed up dataset assembly for users with access to high performance compute clusters. The use of the PostgreSQL Relational Database Management System is a good initial step towards this goal, since it is designed to allow multiple simultaneous connections.

Taxonomy storage

Currently, TaxMan stores only a single reference taxonomy, that provided by the NCBI (web reference 6). This is because TaxMan uses sequence data obtained from the GenBank database, and the NCBI taxonomy is the only scheme that includes all of the species to which records are assigned. A useful addition to TaxMan would be the ability to store multiple reference taxonomies for a set of taxa of interest. Because both taxonomies and phylogenetic trees are stored internally in TaxMan using a common format, it would be possible to compare them. For instance, code could be added to automatically evaluate congruence between a phylogenetic tree and two competing taxonomic schemes.

Taxon selection

Currently, the only criterion on which TaxMan can select species to represent higher taxa is sequence completeness. Although this is a large component of suitability of a species for phylogenetic analysis, it should not be the only consideration. Factors such as base composition and evolutionary rate should also be taken into consideration.

2.7.6 Conclusions

There are many scenarios in which researchers might want to assemble a dataset of aligned DNA or protein sequences with a view to phylogenetic analysis. Current approaches require assembling such datasets manually or, at best, using software tools that lack integration. TaxMan greatly facilitates the various steps required for a multigene phylogenetic study by extracting sequences from public databases, building and aligning consensus sequences, choosing sets of data for analysis and storing the results of analysis. By using a relational database to store data, along with existing bioinformatics tools and a Perl framework, TaxMan allows large datasets to be assembled extremely rapidly and with full data provenance, allowing the user to concentrate on the analysis.

2.8 Technical notes

2.8.1 Current release version and date

TaxMan 1.1_rc1 05/09/2006

2.8.2 Availability

TaxMan is available from

`www.nematodes.org/bioinformatics/TaxMan/`

under the GNU General Public Licence

2.8.3 Dependencies

Linux (developed on Fedora Core 4, 2.6 kernel series)

PostgreSQL (developed on version 8.0.8)

Perl (developed on version 5.8.8)

BioPerl (developed on version 1.5)

Perl modules

- DBI
- DBD::Pg
- Term::ANSIColor

Bioinformatics tools

- BLAST
- EMBOSS
- POA
- phrap

2.8.4 User Guide

A user guide is provided as Appendix 1.

3 Phylogenetic analysis of Chelicerates using mitochondrial genes

3.1 Abstract

Chelicerates are a diverse group of arthropods which are important in the study of several biological phenomena. A robust phylogeny for the chelicerate orders would facilitate comparison of characters across chelicerates as well as informing the pattern of arthropod evolution. In this chapter I describe the use of multiple mitochondrial genes to construct a phylogeny of the chelicerate orders using Bayesian analyses, with particular emphasis on the position of scorpions. I find that the phylogeny is extremely sensitive to the evolutionary model used, specifically with respect to the partitioning scheme, and to the use of a recoding scheme designed to counter the effects of mitochondrial strand-bias. Under the best available model, scorpions are sister taxon to all other arachnids with robust support. Under inferior models, however, alternative hypotheses are given equally robust support. I conclude that mitochondrial genes can be positively misleading under incorrect models and that mitochondrial strand bias renders many previous studies questionable. I find that some taxa with a large amount of missing data can be robustly placed in phylogenetic analysis; that mitochondrial genes place pycnogonids in an untenable position, and that nuclear ribosomal RNA genes are unable to resolve ordinal-level relationships.

Some of the material in this chapter has been written up into a paper, accepted for publication in Molecular Phylogenetics and Evolution (see Appendix 3). Martin Jones

carried out the data gathering and analysis and wrote the paper. Mark Blaxter assisted with interpretation of the results and supervised the project. Benjamin Gantenbein and Victor Fett sequenced the *Mesobuthus gibbosus* mitochondrial genome. All authors assisted with the drafting of the paper.

3.2 Introduction

3.2.1 Chelicerate phylogenetics

The chelicerates (subphylum: Chelicerata) are a diverse group of arthropods characterised by the presence of chelicerae, the first pair of appendages found on the prosoma. Chelicerata comprises Arachnida (containing spiders, scorpions, ticks and mites and several less well-studied groups) and Xiphosura (Horseshoe crabs). Pycnogonida (sea spiders) is usually considered a sister taxon to Chelicerata (Manuel *et al.* 2006; Wheeler and Hayashi 1998) and the two are often united under the name Cheliceriformes. Additionally, the prominent fossil group Eurypterida (water scorpions) are included in Chelicerata under most schemes (Sutton *et al.* 2002).

Several factors motivate the study of chelicerate relationships.

- Spiders (order:Araneae) are a highly speciose group that have been the subject of investigation into the evolution of complex behaviour (Blackledge and Gillespie 2004) and of the genetics of segmental development (Damen, Janssen and Prpic 2005; Schoppmeier and Damen 2005).
- Ticks and mites (order: Acari) contains many species of economic and medical importance, both as parasites in their own right (e.g. *Riphicephalus microplus*,

a cattle tick) and as vectors of disease (e.g. *Ixodes scapularis*, a Lyme disease vector).

- Horseshoe crabs (order: Xiphosura) have been described as 'living fossils', species whose morphology has remained apparently unchanged for a long period of evolutionary time (~250 million years). As such, they are of interest in the study of morphological evolution, which can only be conducted in the context of their relationships to other extant chelicerates.
- Sea spiders (Pycnogonids) are enigmatic, stricly marine animals with several unique features (the proboscis, the ovigers, and a striking body form). Their affiliation with chelicerates is strongly supported (Regier and Shultz 2001; Siveter *et al.* 2004), but their specific place in the arthropod phylogeny is debated (Hassanin 2006).

As one of the extant subphyla of the arthropod phylum, along with Crustacea, Hexapoda and Myriapoda, the relationships between chelicerates and other arthropods is also the subject of speculation and study. Several hypotheses regarding the relationships between these four groups have been proposed, along with the possibility that Crustacea and Hexapoda may be mutually paraphyletic (Cook, Yue and Akam 2005; Mallatt and Giribet 2006; Mallatt, Garey and Shultz 2004; Nardi *et al.* 2003; Regier and Shultz 2001; Regier, Shultz and Kambic 2005).

The aim of this work was to investigate the potential for multi-gene phylogenetics to answer questions regarding the ordinal-level phylogeny of the chelicerates. Relationships between chelicerate orders remain disputed, and many groups are assigned different ranks under different schemes. Table 3.1 lists the orders analysed in

this chapter. This taxonomy was derived from the NCBI taxonomy for each species with relevant sequence data in the GenBank dataset and is not intended to favour any phylogenetic hypotheses

Class	Subclass	Order	Common name
Pycnogonida		Pantopoda	sea spiders
Merostomata		Xiphosura	Horseshoe crabs
Arachnida	Acari	Astigmata	mites and ticks
		Endeostigmata	mites and ticks
		Holothyrida	mites and ticks
		Ixodida	ticks
		Mesostigmata	mites and ticks
		Opilioacarida	mitest and ticks
		Oribatida	beetle mites
		Trombidiformes	mites
		Amblypygi	whip spiders
		Araneae	spiders
		Opiliones	harvestmen
		Palpigradi	microwhip scorpions
		Pseudoscorpiones	false scorpions
		Ricinulei	hooded tickspiders
		Scorpiones	scorpions
		Solifugae	sun spiders
		Uropygi	whip scorpions

Table 3.1: Chelicerate ordinal names used in this chapter

Four competing hypotheses of chelicerate relationships based on morphological character analyses are current. Three have a monophyletic Arachnida (scorpions, spiders, ticks, mites and allies) but differ in the placement of Scorpiones: (1) as a sister group to all other arachnids (reviewed in Wheeler and Hayashi 1998) *versus* (2) Scorpiones as derived arachnids as part of the Dromopoda (including harvestmen and

allies) (Schultz 1990; Wheeler and Hayashi 1998) *versus* (3) Scorpiones as a sister group to Araneae (spiders) rather than Acari (ticks and mites) (Dávila *et al.* 2005; Giribet *et al.* 2002). The fourth hypothesis suggests that (4) scorpions are sister group of the extinct Eurypterida, and thus that arachnids are paraphyletic and that scorpions are the only extant sister group to all other extant chelicerates (Dunlop and Braddy 2001). Although Chelicerata (Arachnida + Merostomata), Acari and Arachnida are often assumed to be monophyletic, the branching order within Arachnida is largely unresolved, with many relationships having been proposed between ticks and mites, spiders, scorpions and the several less well-studied orders. The placement of Pycnogonida (sea spiders) within the chelicerates also varies in different analyses (Hassanin 2006; Wheeler, Giribet and Edgecombe 2004).

Scorpiones as a sister group to other arachnids (hypothesis 1, Figure 3.1) is supported by a number of morphological characters and represents the more traditional view (Wheeler and Hayashi 1998). The Dromopoda hypothesis (Hypothesis 2) was proposed by Schultz (1990) based on parsimony analysis of mostly skeletomuscular characters and supported by Wheeler and Hayashi (1998) with an expanded morphological dataset, and also molecular evidence from nuclear small subunit (18S) and large subunit (28S) rDNA. Scorpions as sister group of spiders as opposed to Acari (hypothesis 3) was reported by Dávila *et al.* (2005) from analysis of mitochondrial genomes, and supported by Giribet *et al.* (2005) using nine gene loci combined with morphology. Dunlop and Braddy (2001) included fossil chelicerate taxa and analysed 33 morphological characters and found support for the Dromopoda. However, Dunlop and Braddy (2001) argued that the morphological dataset is weighted in favour of

skeletomuscular characters; when their analysis was limited to characters apparent only in fossil taxa, it supported a sister relationship between the extinct Eurypterida and Scorpiones, rendering Arachnida paraphyletic (hypothesis 4). Molecular analysis based on six mitochondrial protein-coding genes by Hassanin (2006) seems to support this, as scorpions appear as a sister group to all other chelicerates, and Arachnida are paraphyletic (hypothesis 4).

Each of these four hypotheses have important consequences for the study of chelicerate evolution. If Scorpiones is sister taxon to all other chelicerates (hypothesis 4), shared characters common to scorpions and other chelicerate groups probably represent the ancestral state, and allow us to infer character loss in groups lacking them. Additionally, characters previously though to be synapomorphic for Arachnida are rendered symplesiomorphic, as "Arachnida" is paraphyletic. This is not the case if Scorpiones is sister taxon to Araneae (hypothesis 3), in which case characters shared by scorpions and spiders may be synapomorphies. Similarly, under the Dromopoda hypothesis (hypothesis 2) characters shared by scorpions and other dromopods may be interpreted as synapomorphies. If Scorpions is sister taxon to all other arachnids (hypothesis 1), characters shared by scorpions and any other arachnid group may be interpreted as synapomorphies.

Although scorpions are a key taxon for the understanding of chelicerate evolution, current DNA sequence datasets are limited in extent. In particular, scorpion mitochondrial genomes have not been fully exploited (Gantenbein and Largiadèr 2003). Fourteen complete mitochondrial genomes from Acari (ticks and mites), three from spiders, one from a New World scorpion (*Centruroides limpidus:* Buthidae), and

one from the xiphosuran *Limulus polyphemus* have been sequenced. Additionally, collaborator Benjamin Gantenbein has sequenced the mitochondrial genome of the scorpion *Mesobuthus gibbosus*. In this chapter I describe the use of chelicerate mitochondrial sequence data to explore chelicerate relationships. While these sequence data are unable to differentiate between hypotheses 1 and 2 above (due to lack of data from other dromopods) I can test relative support for hypotheses 1/2 (Figure 3.1 a), *versus* hypothesis 3 (Figure 3.1 b) or hypothesis 4 (Figure 3.1 c).

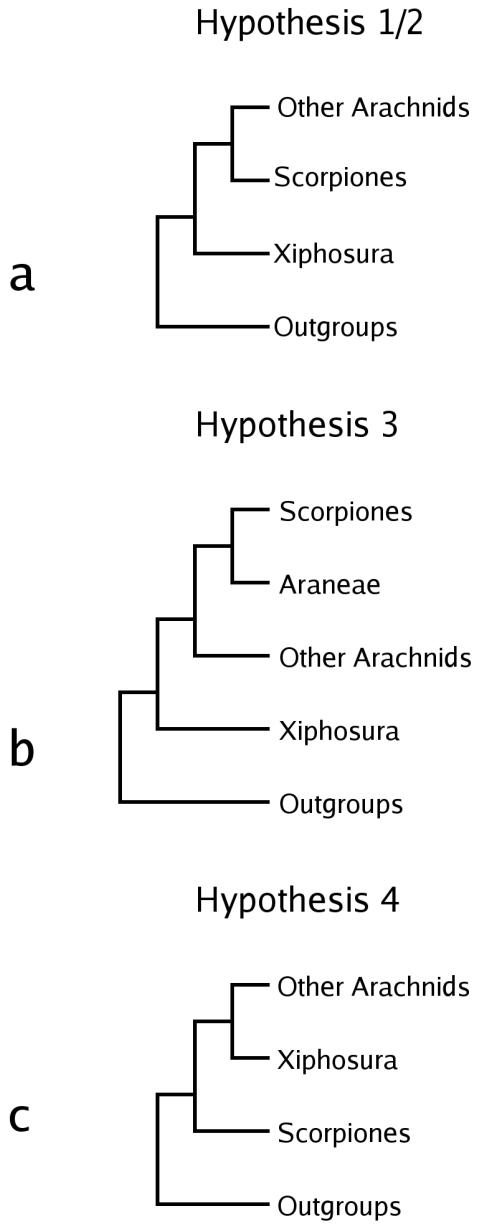


Figure 3.1: Hypotheses regarding the phylogenetic position of scorpions

3.2.2 Multigene studies

There have been relatively few previous attempts to clarify chelicerate relationships using multiple genes. Wheeler and Hayashi (1998) used combined 18S and 28S ribosomal RNA sequence data, along with morphological characters, to investigate relationships between a range of chelicerate taxa. The authors used parsimony and a method that reconstructs phylogenies from unaligned nucleotide sequences, investigating the effects of different weighting schemes (Janies and Wheeler 2002). They concluded that both Chelicerata (Xiphosura + Arachnida) and Arachnida were strongly supported, and placed Scorpiones in a clade with Solifugae, Pseudoscorpiones and Opiliones.

Hassanin (2006) investigated relationships between a large number of arthropod species, including representatives from Crustacea, Insecta, Myriapoda and Chelicerata. The author investigated the effects of (1) mitochondrial strand-bias (see Section 3.2.3) and (2) long branches on the reconstruction of arthropod relationships using Maximum Likelihood (ML) and Bayesian methods. He found that the mitochondrial genomes of some arthropod groups, including spiders and scorpions, were characterised by a reversed mitochondrial strand-bias, causing them to group together under phylogenetic reconstruction. By recoding nucleotide sequences using a scheme designed to eliminate the effects of strand-bias, and excluding taxa with long branches, Hassanin generated an arthropod phylogeny in which the artefactual clustering of reverse strand-bias taxa was not present. Surprisingly, this phylogeny robustly placed Scorpiones as sister taxon to all other chelicerates, resulting in a clade (other Arachnida + Xiphosura) and rendering Arachnida paraphyletic (Hypothesis 4, Figure 3.1 c).

Mallatt and Giribet (2006) used combined 18S and 28S ribosomal RNA genes to evaluate arthropod phylogenetics. Although the description of chelicerate relationships was not the aim of their study, the phylogeny supported a (Xiphosura + Araneae) clade to the exclusion of Scorpiones, again rendering Arachnida paraphyletic (albeit with only three chelicerate taxa represented).

Here, I examine the utility of mitochondrial sequence data for addressing questions of chelicerate phylogeny. Recent data suggest that large datasets, comprising many genes, can resolve problematic phylogenies with a high degree of confidence (Hassanin 2006; Philippe, Lartillot and Brinkmann 2005; Rokas *et al.* 2003). By combining the information from multiple gene sequences, clades can be recovered that are not recovered under analysis of any of the individual genes. Different genes with different evolutionary rates may give strong phylogenetic signals at different depths in a phylogenetic tree. Thus, by including multiple genes in a phylogenetic study, one could obtain a tree that any single gene would be unable to resolve. The use of multiple genes for phylogenetics comes with its own set of difficulties, the most significant of which are computational complexity, and the need for evolutionary models that describe the variation between genes . As the number of genes (and hence the number of characters) included in a multiple sequence alignment grows, so does the time required to evaluate the likelihood or parsimony score of a corresponding phylogenetic tree and hence the time required to execute tree search algorithms. The choice of evolutionary model, always a critical issue in phylogenetic reconstruction , is particularly important where multiple genes are involved (Pupko *et al.* 2002). If the genes evolve under different evolutionary constraints, a single model of DNA

evolution may not be accurately describe the history of all characters in the alignment, and separate models and parameters may have to be assigned to each gene. Additionally, if some gene sequences are unavailable for some taxa, the alignment may have an appreciable proportion of missing data which may adversely affect the robustness of the tree (Wiens 2003).

3.2.3 Strand-bias

The promise of large-scale phylogenetic studies involving multiple genes is that, by combining the phylogenetic signal from a large amount of sequence data, they will be able to resolve relationships that would not be apparent from analysis of single genes. Additionally, multiple-gene phylogenetics offers the potential to combine signal from genes which are informative at different phylogenetic levels, potentially leading to a tree with good resolution from the root to the tips. A quickly-evolving gene might offer good resolution of recent events, but be agnostic regarding ancient events due to mutational saturation. A slowly-evolving gene might contain phylogenetic signal about ancient events, and thus offer resolution near the base of the tree, but be too well-conserved near the tips to contribute any information. A collection of concatenated multiple sequence alignments (a supermatrix) that contained both types of genes would contain phylogenetic information at all levels, and could potentially be used to construct a robust and resolved tree that either gene alone could not have produced.

Essentially, using multiple genes for phylogenetics overcomes inaccuracy due to

limited amounts of phylogentic signal. Assuming that all selected genes are orthologous and share the same evolutionary history, and are correctly aligned, adding more genes to a supermatrix will always increase the amount of phylogenetic signal present. A potential problem in using multiple genes to reconstruct phylogeny lies in the existence of systematic biases that affect entire genomes. Whereas in single-gene studies *stochastic* bias (leading to random phylogenetic signal, or 'noise') is a major factor limiting phylogenetic accuracy, in multiple gene studies systematic bias is likely to be limiting.

Systematic bias can be defined as any characteristic of a genome which tends to affect all genes. Several examples of systematic biases have been found to have disruptive effects on phylogenetic reconstruction. Differing substitution rates between lineages can lead to widely varying branch lengths and the well-known phenomenon of long branch attraction (Brinkmann *et al.* 2005; Felsenstein 1978). Heterotachy – differing patterns of substitution rates between lineages – can also lead to phylogenetic artefacts (Gadagkar and Kumar 2005; Philippe *et al.* 2005). Compositional bias that affects entire genomes, such as AT content, is known to be problematic in the context of likelihood models that assume equal base frequencies across taxa (Galtier and Gouy 1995). Model choice has been shown to be a key factor in overcoming difficulties associated with analysis of biased sequences (Lemmon and Moriarty 2004; Posada and Buckley 2004; Sullivan and Swofford 2001). Additionally, differing patterns of evolution between genes can cause problems in phylogenetic reconstruction under models that fail to take inter-gene differences into account (Nylander *et al.* 2004).

Strand-bias is a particular from of compositional bias which is characteristic of

mitochondrial genomes. In most mitochondrial genomes thus studied, there is an asymmetry in base composition between strands, with one strand having negative CG and AT skew (Formula 3.1) - an abundance of G over C and of T over A (Hassanin, Leger and Deutsch 2005).

$$CG\,skew = \frac{C-G}{C+G}$$

$$AT\,skew = \frac{A-T}{A+T}$$

Formula 3.1: CG and AT skew calculation

The G- and T-rich strand is designated 'heavy' (H) while the G- and T-poor strand is designated 'light' (L) owing to their different buoyancies in a density gradient. Since the strand-bias effect is most pronounced at fourfold-degenerate sites and intergenic regions, it is thought to be the result of substitution bias rather than selection (Hassanin, Leger and Deutsch 2005). A plausible mechanism to explain the substitution bias is the asymmetric nature of mitochondrial genome replication, in which the H strand spends longer in the single-stranded state than the L strand (Clayton 1982; Tanaka and Ozawa 1994). Coupled with the absence of protective histones in mitochondrial DNA, it is thought that this increased time in the single-stranded state could lead to an increased rate of deamination of C and A bases. Deamination of C into U on the H strand would lead to complementary pairing with A on the L strand, with U consequently replaced with T, the overall change being [C->T]

on the H strand and [G->A] on the light strand (Figure 3.2)

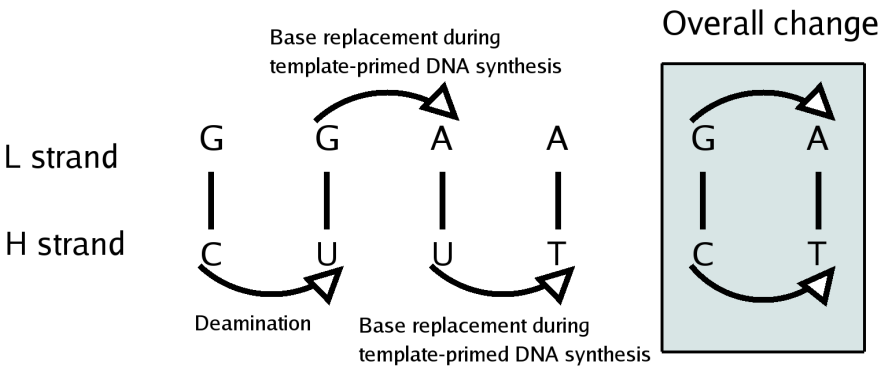


Figure 3.2: Effect of a C->U deamination on the H strand

Similarly, deamination of A to hX (hypoxanthine) on the H strand leads to complementary replacement of T by C on the L strand and an overall change of [T->C] on the L strand and [A->G] on the H strand (Figure 3.3).

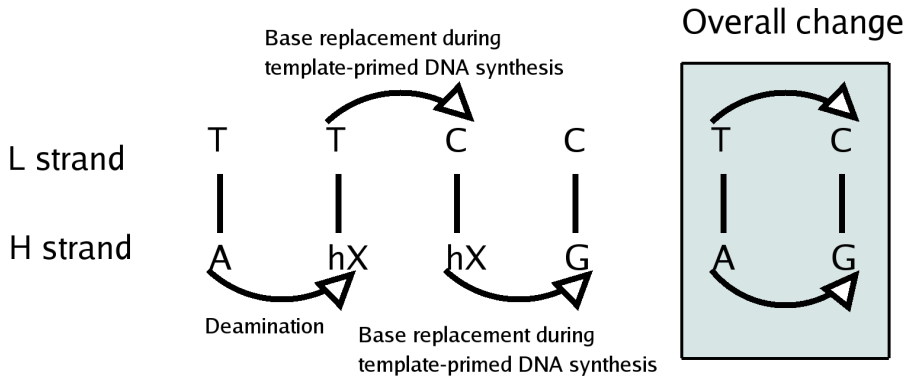


Figure 3.3: Effect of a A->hX deaminatinon on the H strand

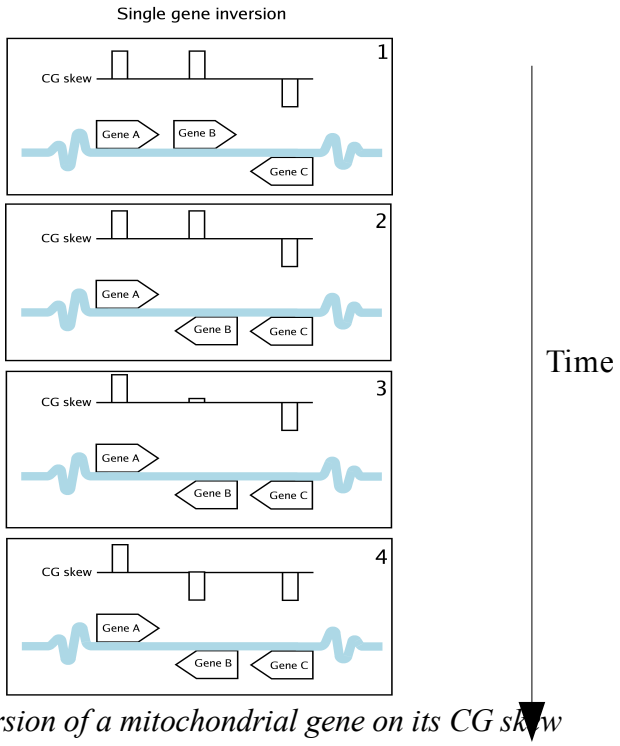


Figure 3.4: Effect of inversion of a mitochondrial gene on its CG skew

Each panel shows a region of a mitochondrial genome containing three genes and, above, a bar chart showing the CG skew for each gene.

Panel 1 - Genes A and B are on the top strand and have positive CG skew, Gene C is on the bottom strand and has negative CG skew (measured in 5′ – 3′ coding direction).

Panel 2 - Gene B has undergone an inversion and now occupies the bottom strand. Immediately after inversion, its CG skew is unchanged.

Panel 3 - The situation some time after the inversion event. Strand bias has acted to decrease the CG skew of Gene B, which is changing from positive to negative.

Panel 4 - The situation long after the inversion event. Gene B's CG skew is negative, like that of the other genes on the bottom strand, and is now stationary. Gene B in this lineage now has the opposite CG skew to gene B in other lineages in which the inversion event has not taken place (in which the situation resembles panel 1) and to other genes which were formerly on the same strand (Gene A). If gene B undergoes independent inversion events in different taxa, those taxa will cluster together under phylogenetic analysis due to convergent changes caused by strand bias

Because the strand-bias effect affects all regions of the mitochondrial genome, protein coding genes will acquire 5' to 3' CG and AT skews corresponding to the strand on which they are located. Because protein-coding genes can be on either strand of a mitochondrial genome, different genes will exhibit contrasting 5' – 3' skews. There are two scenarios in which the skew for a gene can be reversed. If a mitochondrial gene undergoes an inversion in a particular lineage it will experience a reversed substitution bias and will start accumulating the opposite strand-bias. Alternatively, if the mitochondrial control region, which controls the polarity of mitochondrial genome replication, undergoes an inversion event, the roles of the two strands during replication will be reversed, so that genes formerly on the H strand will begin to accumulate L strand bias. In the first scenario, the inverted gene will acquire opposite skews to other mitochondrial genes that were formerly on the same strand, and to the same mitochondrial gene in closely related taxa (Figure 3.4). In the second, all mitochondrial genes will acquire opposite skews – the pattern of skew along the genome as a whole will have been reversed (Figure 3.5). In this case, each mitochondrial gene will acquire the opposite skew of the same gene in closely related taxa. However, the process will be dynamic, with skew gradually changing due to accumulated substitutions following the inversion event.

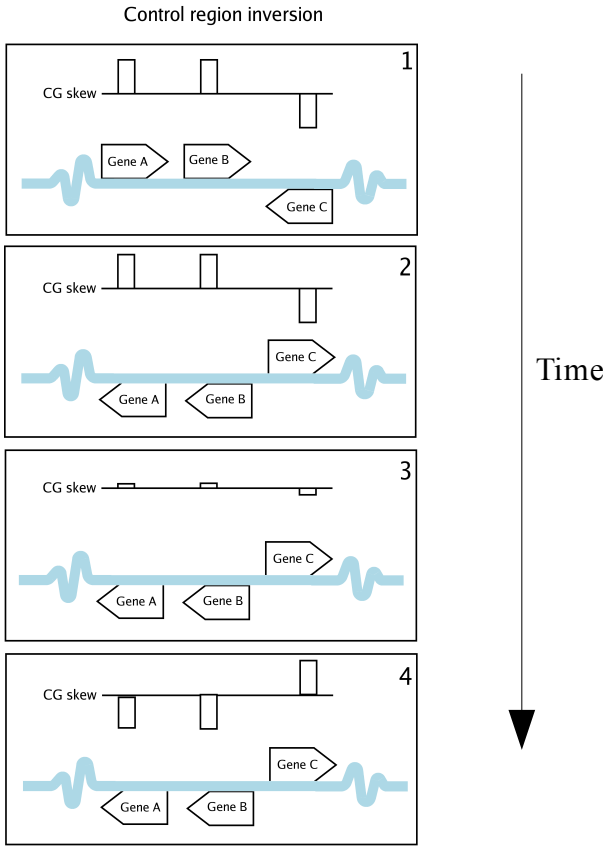


Figure 3.5: Effect of mitochondrial control region inversion on CG skew

Each panel shows a region of a mitochondrial genome containing three genes and, above, a bar chart showing the CG skew for each gene.

Panel 1 - Genes A and B are on the top strand and have positive CG skew, Gene C is on the bottom strand and has negative CG skew.

Panel 2 - An inversion of the mitochondrial control region has reversed the polarity of mitochondrial genome replication, effectively causing the strands to swap roles. The genes are shown as inverted relative to panel 1. CG skew is initially unchanged.

Panel 3 - The situation some time after the inversion of the mitochondrial control region. Strand bias has acted to reverse the CG skew of all three genes. For Genes A and B, the CG skew is changing from positive to negative, whereas for Gene C the CG skew is changing from negative to positive.

Panel 4 - The situation long after the inversion of the mitochondrial control region. Each gene's CG skew is stationary and reversed relative to that before the mitochondrial control region inversion and that in lineages where no inversion has taken place (Panel 1).

The consequences for phylogenetic analysis are clear. If sequences with both types of strand-bias are included in a multiple sequence alignment, any single model of sequence evolution will be unable to accurately describe the history of the sequences. The two groups of sequences will differ in terms of nucleotide frequencies; also, the appropriate substitution rate matrix will be different for different parts of the tree.

In phylogenetic studies, taxa with reverse patterns of strand-bias relative to the majority cluster together artefactually when analysed under standard evolutionary models (Hassanin 2006). Hassanin specified a novel evolutionary model (Neutral Transitions Excluded, NTE) which aims to remove the effects of strand bias by recoding a proportion of nucleotides as purines (R) and pyrimidines (Y). Recoding is applied to positions at which a transition would result in a neutral or "nearly neutral" substitution (Table 3.2).

Before	After	Before	After	Before	After	Before	After
AAA	AAR	GAA	GAR	TAA	TAR	CAA	CAR
AAG	AAR	GAG	GAR	TAG	TAR	CAG	CAR
AAT	AAY	GAT	GAY	TAT	TAY	CAT	CAY
AAC	AAY	GAC	GAY	TAC	TAY	CAC	CAY
AGA	AGR	GGA	GGR	TGA	TGR	CGA	CGR
AGG	AGR	GGG	GGR	TGG	TGR	CGG	CGR
AGT	AGY	GGT	GGY	TGT	TGY	CGT	CGY
AGC	AGY	GGC	GGY	TGC	TGY	CGC	CGY
ATA	RYR	GTA	RYR	TTA	YTR	CTA	YTR
ATG	RYR	GTG	RYR	TTG	YTR	CTG	YTR
ATT	RYY	GTT	RYY	TTT	YTY	CTT	YTY
ATC	RYY	GTC	RYY	TTC	YTY	CTC	YTY
ACA	RYR	GCA	RYR	TCA	TCR	CCA	CCR
ACG	RYR	GCG	RYR	TCG	TCR	CCG	CCR
ACT	RYY	GCT	RYY	TCT	TCY	CCT	CCY
ACC	RYY	GCC	RYY	TCC	TCY	CCC	CCY

Table 3.2: NTE recoding scheme

Each large block contains two columns which show the effect of NTE recoding on a single codon.

Neutral substitutions do not change the amino acid for which a codon codes, and can come about due to the redundancy of the genetic code. "Nearly neutral" substitutions are those that do change the amino acid for which a codon codes, but in which the replacement amino acid has very similar chemical properties to the old one. Positions in which a transition would be neutral or nearly neutral are those in which substitutions are most likely to be driven by strand bias; by recoding them we can attempt to eliminate the effect of strand bias while retaining genuine phylogenetic signal. When analysed under the NTE model, a dataset of six mitochondrial protein-coding genes for 71 arthropod species yielded a phylogeny with Chelicerata monophyletic and Scorpiones as sister taxon to all other chelicerates (Hassanin 2006 Hypothesis 4, Fig. 3.1). To investigate the robustness of this result, and the utility of the NTE model for chelicerate phylogeny using mitochondrial data, I used a bioinformatics pipeline, TaxMan (see Chapter 2), to mine publicly available sequence data, including the recently-sequenced mitochondrial genome of *M. gibbosus* from collaborator B. Gantenbein, and assemble a dataset of aligned chelicerate mitochondrial genes, including species for which only a few mitochondrial gene sequences were available. Bayesian phylogenetic analysis was performed on subsets of these sequences using a variety of evolutionary models. Additionally, I reanalysed the dataset used in Wheeler and Hayashi (1998), hereafter the W1 dataset.

3.3 Methods

3.3.1 TaxMan

The TaxMan software package (described in detail in Chapter 2) was used to assemble a dataset of aligned mitochondrial genes for the subphylum Chelicerata. For a detailed discussion of sequence extraction, consensus building and alignment of this dataset, see Section 2.6.1

3.3.2 Phylogenetic Analysis

Multiple sequence alignments suitable for phylogenetic analysis were generated by specifying subsets of genes and taxa (slices) and extracting the corresponding pre-aligned sequences from the database (Table 3.3). The alignments were analysed using MrBayes 3.1 (Ronquist and Huelsenbeck 2003). Several evolutionary models were used in the analysis of the D1 and D2 datasets (Table 3.4). By default, a GTR model (nst=6) was applied to all alignments. For those models where the NTE scheme was used, bases were recoded according to Hassanin, Leger and Deutsch (2005) and third codon position bases were assigned to a separate partition with a two substitution types (nst=2). For those models where the alignment was partitioned by gene, base frequencies, substitution rates, alpha parameters and proportions of invariant sites were unlinked across partitions, and a rate multiplier was used to allow rate variation between partitions. For analysis of the W1 dataset, a GTR (nst=6) model was applied with gamma rate variation and a proportion of invariant sites. No partitioning was

carried out in the W1 analyses.

Taxon	Order[1]	NCBI TXID [2]	RefSeq ID[3]	Citation[4]	D1 (12669 characters) [5]		D2 (15950 characters) [6]	
					no. chars [7]	% present [8]	no. chars [7]	% present [8]
Unidentified Opilioacarid	Opilioacarida	150113	-	-	-	-	257	1.61%
Unidentified Allothyrid	Holothyrida	91335	-	-	-	-	539	3.38%
Steganacarus magnus	Oribatida	52000	-	-	-	-	387	2.43%
Camisia horrida	Oribatida	240610	-	-	-	-	597	3.74%
Sarcoptes scabei	Astigmata	197185	-	-	-	-	7394	46.36%
Opilio parietinus	Opiliones	121214	-	-	-	-	478	3.00%
Ixodes hexagonus	Ixodida	34612	NC_002010	Black and Roehrdanz, 1998	11820	93.30%	14449	90.59%
Mastigoproctus giganteus	Uropygi	58767	-	-	3756	29.65%	4150	26.02%
Limulus polyphemus	Xiphosura	6850	NC_003057	Lavrov, Boore and Brown, 2000	12108	95.57%	14692	92.11%
Argiope bruennichi	Araneae	94029	-	-	4857	38.34%	4863	30.49%
Phrynus sp.	Amblypygi	309714	-	-	5154	40.68%	5165	32.38%
Endeis spinosa	Pycnogonida	136194	-	-	5019	39.62%	5030	31.54%
Mesobuthus gibbosus	Scorpiones	123226	NC_006515	This work	12090	95.43%	14331	89.85%
Euscorpius flavicaudis	Scorpiones	100976	-	-	4926	38.88%	6100	38.24%
Leptotrombidium pallidum	Trombidiformes	279272	NC_007177	Shao et al. 2005	11487	90.67%	13776	86.37%
Heptathela hangzhouensis	Araneae	216259	NC_005924	Qui et al. 2005	11997	94.70%	14472	90.73%
Carios capensis	Ixodida	176285	NC_005291	-	11922	94.10%	14444	90.56%
Scutigera coleoptrata [9]	Scutigeromorpha	29022	NC_005870	Negrisolo, Minelli and Valle, 2004	11979	94.55%	14496	90.89%
Triops cancriformis [9]	Notostraca	194544	NC_004465	Umestu et al. 2002	12063	95.21%	14613	91.62%
Daphnia pulex [9]	Diplostraca	6669	NC_000844	Crease 1999	12135	95.78%	14727	92.33%
Triatoma dimidiata [9]	Hemiptera	72491	NC_002609	Dotson and Beard 2001	12003	94.74%	14578	91.40%
Drosophila melanogaster [9]	Diptera	7227	NC_001709		12123	95.69%	14723	92.31%

Table 3.3: Summary of taxa included in the D1 and D2 datasets

1 – The order to which each species belongs (NCBI taxonomy)
2 – The Taxonomy ID (TXID) assigned to the species by the NCBI GenBank database
3 – The RefSeq ID (where applicable) for the complete mitochondrial genome sequence
4 – The citation (where applicable) for the complete mitochondrial genome sequence
5 – Dataset D1 includes the following genes: ATP6, COX1, COX2, COX3, CYTB, ND1, ND2, ND3, ND4, ND4L, ND5, ND6
6 – Dataset D2 includes the following genes: ATP6, ATP8, COX1, COX2, COX3, CYTB, ND1, ND2, ND3, ND4, ND4L, ND5, ND6, RNA_12S, RNA_16
7 – No. chars: the number of aligned characters in each species
8 – % present: the proportion of bases in the complete alignment present in each species
9 – Outgroup taxa

Analyses were run using default MCMCMC parameters for 1,000,000 generations. Convergence of split frequencies and flattening of likelihood scores were checked to ensure stationarity and the first 100,000 generations (10%) discarded as burn-in. Trees sampled after the burn-in period were summarised to give the 50% majority rule consensus tree. To determine the significance of support for the optimal tree, analyses were run using parameter estimates from the initial analysis with the following constraints: Hypothesis 2 – *Ixodes hexagonus, Mastigoproctus giganteus, Limulus polyphemus, Argiope bruennichi, Phrynus sp., Leptotrombidium pallidum, Heptathela hangzhouensis* and *Cario capensis* monophyletic; Hypothesis 3 – *Argiope bruennichi, Heptathela hangzhouensis, Mesobuthus gibbosus* and *Euscorpius flavicaudis monophyletic.*

3.3.3 CG and AT skew

Using a Perl script, an alignment was produced for each codon position in each gene containing the taxa in the D1 subset. CG and AT skew was calculated for each taxon. For third positions, skews were also calculated following recoding using the NTE scheme, to examine the effects of recoding. Skews were calculated according to Forumulae 3.1 and .

3.4 Results

For results pertaining to the assembly of the dataset, see Section 2.6.1

I examined a subset (D1) of the mitochondrial data consisting of 12 protein-coding genes for species from the chelicerate orders considered in Hassanin (2006). I then examined a second subset (D2) containing sequences from additional chelicerate orders, which incorporated a significant proportion of missing data. Finally, I examined a dataset containing only nuclear ribosomal RNA sequences (W1), and compared the results with those obtained from protein-coding genes. In all analyses of D1 and D2, Pycnogonida was placed within Acari. This result is unexpected, because morphological evidence places pycnogonids distant from ticks and mites (Maxmen *et al.* 2005). While strongly supported, this result may be due to the relatively small amount of sequence data available for Pycnogonida (~40% in D1, ~30% in D2, see Table 3.3) and I have therefore ignored Pycnogonida when describing relationships in the following sections. *Varroa destructor*, the honeybee mite, is an important economic parasite and as such has been well-studied and has copious sequence data available (Navajas *et al.* 2002). However, the extremely high A+T content of its mitochondrial genome renders it unsuitable for phylogenetic analysis and I did not consider it for inclusion in either dataset.

Analysis name[1]	Model of base change	Codon positions	Recoded	Partitioned	D1 tree[7]	D2 tree[7]
GTRp	GTR [2]	1,2	n	gene [5]	4	4
NTE	NTE [3]	1,2	y [4]	n	4	-
NTEp	NTE [3]	1,2	y [4]	gene+3 [6]	4	-
GTR+3p	GTR [2]	1,2,3	n	gene [5]	3	-
NTE+3	NTE [3]	1,2,3	y [4]	n	4	-
NTE+3p	NTE [3]	1,2,3	y [4]	gene+3 [6]	1/2	1/2

Table 3.4: Details of analysis of the D1 and D2 datasets

1 – Name used to refer the analysis in the text. All analyses include gamma rate variation and a proportion of invariant sites. Codon positions 1 and 2 are included in all analyses. Inclusion of 3rd codon position bases in the latter three analyses is indicated by '+3'. Analyses that are partitioned by gene end in p.

2 – General Time Reversible (GTR) model applied to all bases (nst=6 in MrBayes)

3 – Neutral Transitions Excluded (NTE) model; GTR applied to first and second codon position bases (nst=6) and two substitution type model applied to third codon position bases (nst=2).

4 – Bases are recoded according to the NTE scheme of Hassanin (2006). In practise, all third codon position bases and a subset of first and second codon position bases are recoded as purine/pyrimidine (R/Y), eliminating phylogenetic signal caused by neutral and nearly neutral transitions.

5 – The alignment is partitioned by gene. Each partition has independent base frequencies, transition rate matrix, gamma parameter and proportion of invariant sites. A rate multiplier is used to permit rate variation between partitions.

6 – First and second codon position bases are partitioned by gene and all third codon position bases form a separate partition. Each partition has independent base frequencies, transition rate matrix, gamma parameter and proportion of invariant sites. A rate multiplier is used to permit rate variation between partitions.

7 – The hypothesis (see Introduction, Fig. 3.1) supported by each dataset under the model.

3.4.1 Dataset D1: Mostly complete mitochondrial genomes

The D1 dataset consisted of twelve protein-coding genes for eleven chelicerate species and five outgroup species (Table 3.3). In order to keep the analysis computationally tractable, I used a subset of taxa for which mitochondrial genomes are available, selecting taxa to obtain the widest taxonomic coverage. In preliminary analyses, all orders were found to be monophyletic and inclusion of multiple representatives of each order did not alter the results. In order to investigate the effect of model choice on the phylogeny, six analyses were carried out (Table 3.4). The analyses differ in (1) the inclusion or exclusion of the third bases of codons (2) the use of the NTE recoding scheme and (3) permitting each gene to have different model parameters (partitioning).

Different phylogenies were obtained from Bayesian analysis of the D1 dataset in the different analyses (Fig. 3.6; hypotheses in Fig. 3.1). The three analyses in which third codon position bases were excluded yielded similar trees (Hypothesis 4) with support for scorpions as sister taxon to other chelicerates (rendering Arachnida paraphyletic). The *GTRp* analysis (Fig. 3.6 a) gave a tree with the remaining arachnids divided into two clades; one of spiders and related orders ((Araneae + Uropygi) + Amblypygi) and one of mites and ticks (Ixodida + Trombidiformes). In the *NTE* analysis (Fig. 3.6 b), an identical tree was recovered with the exception of rearrangements within the two major arachnid clades. Under the *NTEp* analysis (Fig. 3.6 c), the tree was similar to that obtained under the *NTE* analysis except for the relationships within the clade of mites and ticks, which were identical with those of the *GTRp* tree, and the reduced

support for scorpions as sister taxon to other chelicerates (posterior probability of 0.75 compared to 0.95, posterior probabilities of >0.9 are considered significant for Bayesian analysis). When third codon position bases were included, each analysis recovered a different tree. Under the *GTR+3p* model (Fig. 3.6 d) scorpions were no longer sister taxon to all other chelicerates. A clade of scorpions and the spider *Argiope bruennichi* was recovered, with a (Uropygi + *Heptathela hangzhousensis*) clade as a sister group (Hypothesis 3 with the exception of paraphyletic Araneae). A clade of ticks and mites was recovered. A clade consisting of Xiphosura + Amblypygi was a sister taxon to all other chelicerates. Under the *NTE+3* model (Fig. 3.6 e) the tree was identical to that found under *NTE* (Hypothesis 1/2) although with negligible support for scorpions as sister taxon to all other chelicerates. Under the *NTE+3p* analysis (Fig. 3.6 f) Arachnida was monophyletic, with Scorpions a sister taxon to the other studied arachnids, and Xiphosura as sister taxon to all other chelicerates (Hypothesis 1). Two major clades of arachnids were recovered; one of spiders and related orders (Araneae + (Uropygi + Amblypygi) and one of mites and ticks (Ixodida + Trombidiformes).

Analyses under NTE+3p when the tree was constrained to support Hypothesis 3 or Hypothesis 4 yielded harmonic mean log-likelihood estimates of -81,004.98 and -80,996.41 respectively, compared with an estimate of -80,722.24 for the optimal tree. To test whether the data supported use of partitioned model, I estimated harmonic mean log-likelihoods for two pairs of analyses that varied only in the partitioning; *NTE* versus *NTEp* and *NTE+3* versus *NTE+3p.* In each case the partitioned model was preferred: *NTE*: -53885 << *NTEp*: -53193; *NTE+3*: -82056 << *NTE+3p*: -80721.

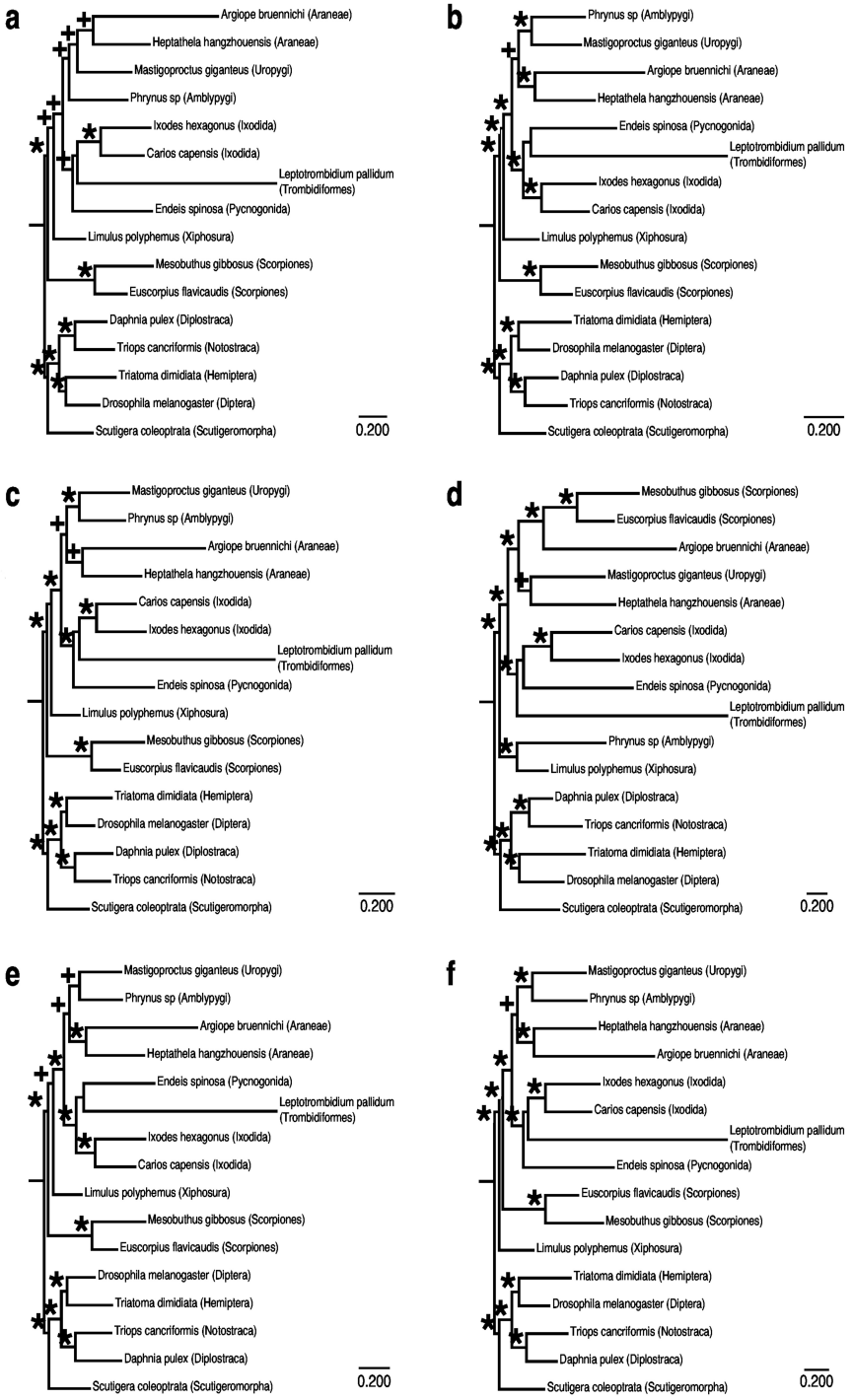


Figure 3.6: Effect of model choice on the phylogeny recovered from the D1 dataset (legend overleaf).

Figure 3.6

Phylogeny reconstructed from the D1 dataset using Bayesian analysis under the following analyses (see Table 3.4):(a) GTRp (b) NTE (c) NTEp (d) GTR+3p (e)NTE+3 (f) NTE+3p. The order to which each species belongs is given in parentheses. The scale bar shows the branch length associated with 0.200 expected changes per site. Labels on branches indicate the level of support: (*) - posterior probability =1.00; (+) - posterior probability = 0.9-0.99. Posterior probabilities of >0.9 are considered significant; lower posterior probabilities are not shown.

3.4.2 Dataset D2: including single mitochondrial genes for some taxa

The D2 dataset consisted of thirteen protein-coding genes and two rRNA genes for seventeen chelicerate species and five outgroup species (Table 3.3). The number and percentage of characters present varied between taxa, ranging from 257 (1.61%) characters present for an unidentified Opilioacarid (Opilioacaridae) to 14723 (92.31%) characters present for *Drosophila melanogaster* (outgroup). Taxa with complete mitochondrial genomes, such as *D. melanogaster* had ~8% missing characters due to minor differences in sequence length between taxa, resulting in end gaps. Missing data is known to be reduce phylogenetic accuracy, but is an inevitable consequence of assembling large alignments from public data. It has been suggested that the problems associated with missing data can be overcome if the absolute number of characters in an alignment is large. I included species with a range of numbers of characters present to investigate the effects of missing data on the phylogenetic placement of 'neglected' taxa.

The results from analysis of D1 indicate that use of the NTE recoding scheme greatly influenced the result of phylogenetic analysis. With this in mind, I analysed D2 under the most comprehensive analysis that used NTE, and under a GTR model for comparison. Under the *NTE+3p* scheme (Fig. 3.7), Xiphosura was robustly placed arising as sister taxon to the remaining chelicerates, with Scorpiones a sister taxon to the remaining arachnids (Hypothesis 1/2). Relationships withing the arachnids were

generally poorly resolved, although some groups were strongly supported: (Ixodida + Holothyrida) and (Uropygi + Amblypygi). Under the *GTRp* scheme Hypothesis 3 was recovered (not shown).

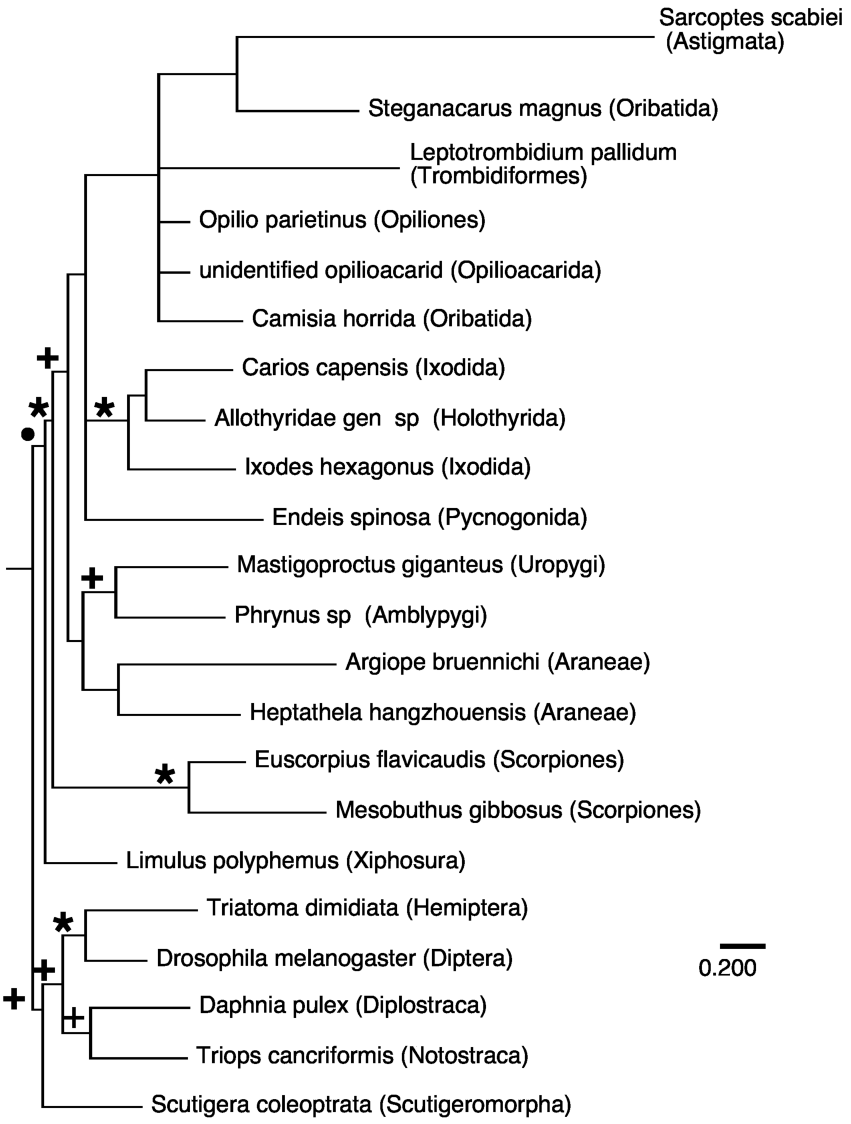


Figure 3.7: Phylogeny recovered from the D2 dataset

Phylogeny reconstructed from the D2 dataset using Bayesian analysis under the NTE+3p scheme (see Table 3.4 for details of models). The order to which each species belongs is given in parentheses. The scale bar shows the branch length associated with 0.200 expected changes per site. Labels on branches indicate the level of support: (*) posterior probability =1.00 , (+) posterior probability = 0.9-0.99, and (•) posterior probability = 0.80-0.89. Posterior probabilities of >0.9 are considered significant. Posterior probabilities of less than 0.80 are not shown.

3.4.3 Dataset W1: A nuclear gene dataset

The *W1* dataset, derived from Wheeler and Hayashi (1998), consisted of the nuclear SSU and LSU RNA genes for twenty three chelicerate species and the same five outgroup species as used in D1 and D2. This dataset formed the basis, along with morphological data, for the summary cladogram of chelicerate order relationships in Wheeler and Hayashi (1998) and Wheeler, Giribet and Edgecombe (2004). Analysis of each gene individually yielded trees that were essentially unresolved. Analysis of a concatenated alignment under a GTR model with gamma rate variation and a proportion of invariant sites yielded a tree that had high levels of support for ordinal-level clades (Scorpiones, Araneae, Uropygi) but which was largely uninformative with regard to relationships between orders. Additionally, it proposed untenable relationships between the outgroup taxa, with Neoptera (Insecta) and Branchiopoda (Crustacea) paraphyletic (Fig. 3.8).

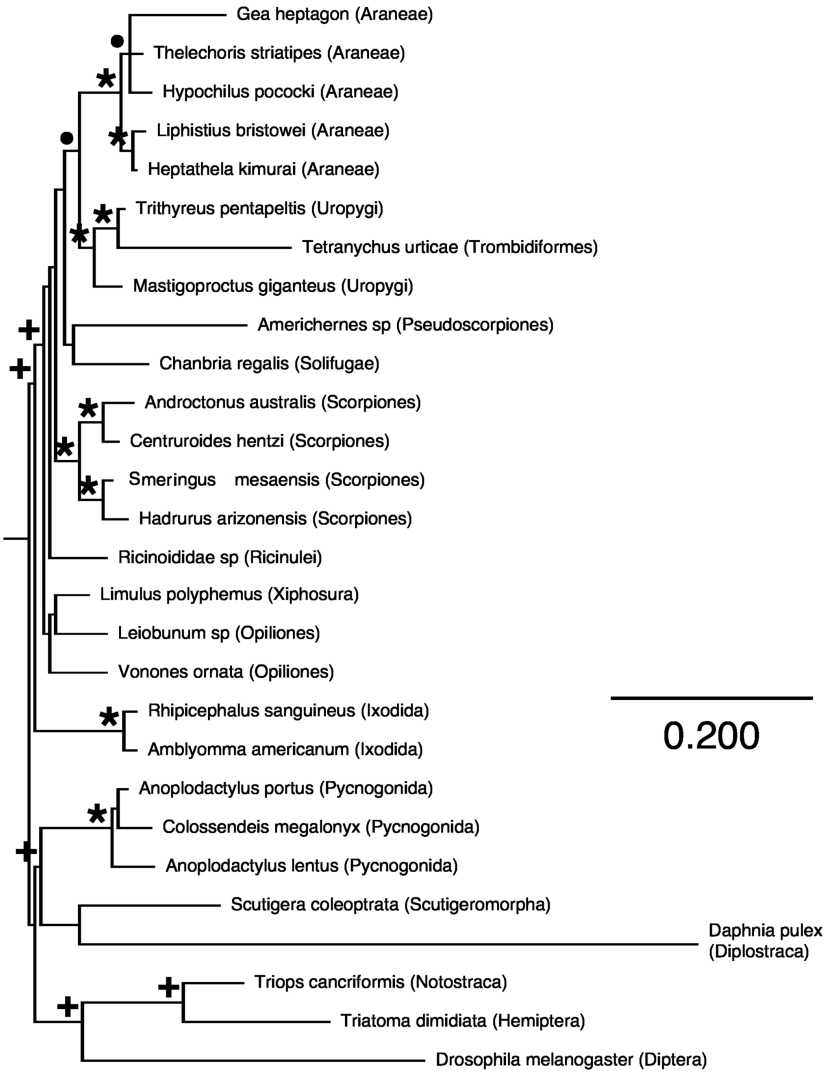


Figure 3.8: Phylogeny recovered from the W1 dataset

Phylogeny reconstructed from the W1 dataset using Bayesian analysis under the GTR model. The order to which each species belongs is given in parentheses. Labels on branches indicate the level of support: (*) posterior probability =1.00, (+) posterior probability = 0.9-0.99, and (•) posterior probability = 0.80-0.89. Posterior probabilities of >0.9 are considered significant. Posterior probabilities of less than 0.80 are not shown. The scale bar shows the branch length associated with 0.200 expected changes per site.

3.4.4 Skew analysis of D1

Figure 3.9 shows the results of CG skew analysis for selected taxa from the D1 subset (plots in the text follow the same format). Several features are apparent. In general, skew is largest in magnitude at third codon position bases and smallest at second codon position bases, with first codon position bases intermediate. Despite this general trend, the pattern is noisy; some genes in some taxa have greater skews at first than second positions and in some taxa first and second positions appear equally skewed (e.g. *L. polyphemus* first  and second  codon position bases – see Figure 3.9 for explanation of minigraphs).

The pattern of skew at third codon position bases is particularly interesting. Most taxa have have a characteristic pattern of positive and negative CG skew (e.g *L. polyphemus*  , *T. cancriformis*  , *H. hangzhouensis* ). This correlates with the position of the coding genes on the mitochondrial genome; genes on opposing strands will have opposing skews. In a few taxa, however, an inverse pattern is seen (*M. gibbosus* , *A. bruennichi* note that some data are missing for this species]). This pattern is best explained by an inversion of the mitochondrial replication origin, leading to reversed mutation pressure on the two strands of the mitochondrial genome.

The NTE recoding scheme removes the majority of CG skew from third position basis,

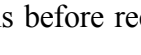
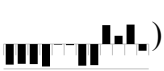
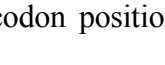
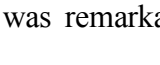
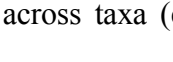
indicating that the skew is mainly present at neutral and nearly neutral sites. Taxa that show opposite skew patterns before recoding (e.g. *T. cancriformis*  vs. *M gibbosus* ) show no strong pattern after recoding ( vs.).

Figure 3.10 shows the corresponding data for AT skews. Unlike CG skews, AT skews did not show a strong, consistent strand effect. In general, skew was weak at first and third codon position bases (e.g. *T. cancriformis* first and third positions). The most striking pattern was strong negative AT skew in second codon position bases which was remarkably consistent across taxa (e.g. *M. gibbosus*  , *L. polyphemus*  , *T. cancriformis* ) .

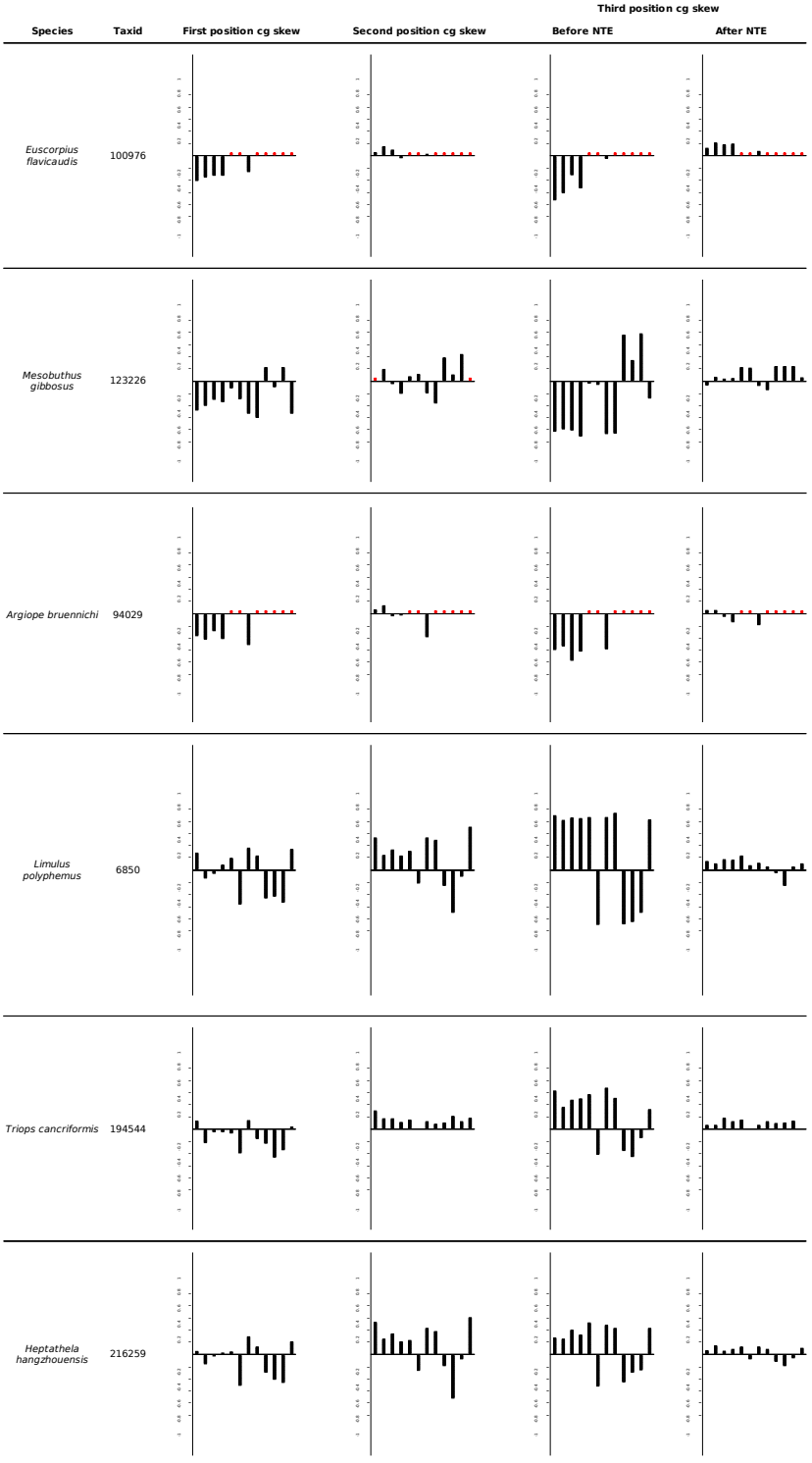


Figure 3.9: CG skew analysis of selected taxa from the D1 dataset (legend overleaf)

Figure 3.9

Each small graph shows the 5′ to 3′ CG skew of each mitochondrial protein-coding gene in alphabetical order, excluding ATP8. Thus, 12 values are shown on each small graph. The scale in each case is 1 to -1, since CG skews can be positive or negative. Missing data are indicated by red X's. Each row shows the data for a particular species, the name and NCBI Taxid being shown in the first and second columns respectively. Four graphs are shown for each species. The first graph shows CG skew measured at first codon position bases. The second graph shows CG skew measured at second codon position bases. The third and fourth graphs show CG measured at third codon position bases before and after the sequences have been recoding using the NTE scheme. The gene order for each graph is as follows:ATP6 COX1 COX2 COX3 CYTB ND1 ND2 ND3 ND4 ND4L ND5 ND6 (ATP8 is excluded due to its short length).

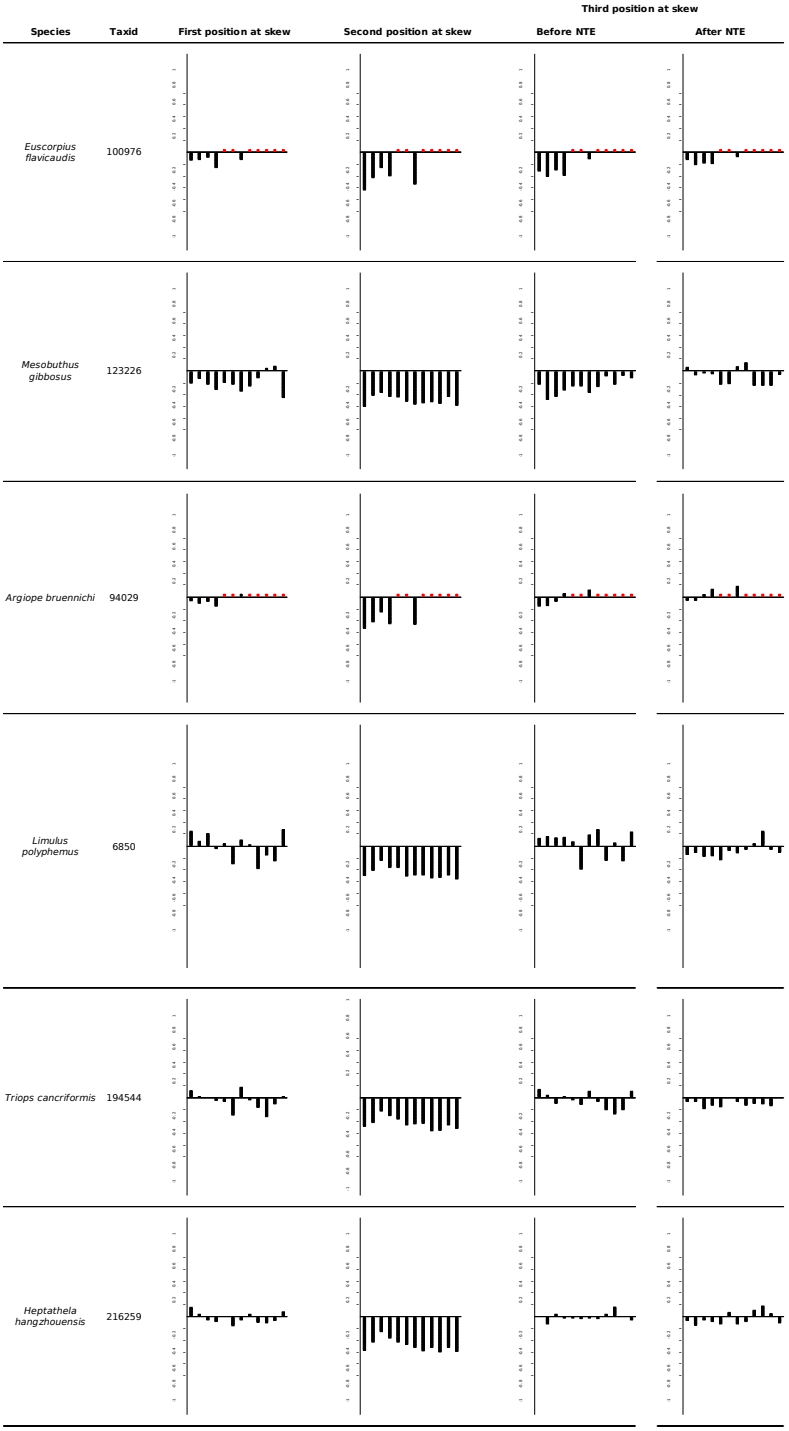


Figure 3.10: Figure 3.10: AT skew analysis of selected taxa from the D1 dataset (legend overleaf)

Figure 3.10

Each small graph shows the 5′ to 3′ AT skew of each mitochondrial protein-coding gene in alphabetical order, excluding ATP8. Thus, 12 values are shown on each small graph. The scale in each case is 1 to -1, since AT skews can be positive or negative. Missing data are indicated by red X's. Each row shows the data for a particular species, the name and NCBI Taxid being shown in the first and second columns respectively. Four graphs are shown for each species. The first graph shows AT skew measured at first codon position bases. The second graph shows AT skew measured at second codon position bases. The third and fourth graphs show AT measured at third codon position bases before and after the sequences have been recoding using the NTE scheme. The gene order for each graph is as follows:ATP6 COX1 COX2 COX3 CYTB ND1 ND2 ND3 ND4 ND4L ND5 ND6 (ATP8 is excluded due to its short length).

3.5 Discussion

3.5.1 Evolutionary models

The reversals of strand-bias documented by Hassanin, Leger and Deutsch (2005) clearly show how systematic error can limit the usefulness of multigene phylogenetic analysis. Multiple independent reversals of strand bias in mitochondrial genomes and individual mitochondrial genes cause erroneous clustering of distantly related arthropod taxa when alignments are subjected to phylogenetic analysis under a standard General Time Reversible (GTR) model. In their work, Hassanin and collegues found evidence that strand bias is mainly generated by transitions (Hassanin 2006; Hassanin, Leger and Deutsch 2005). They proposed a recoding scheme, Neutral Transitions Excluded (NTE) which removes the effect of strand-bias by recoding bases at neutral and nearly-neutral positions as purines and pyrimidines (R/Y coding). Here I have investigated a wider range of evolutionary models than has Hassanin, examining the effects of GTR analysis of the original data compared to NTE analysis of recoded data, the effects of inclusion or exclusion of third codon position bases, and the effects of a partitioned versus an unpartitioned model.

A notable result is the contribution of phylogenetic signal from third codon position bases. Phylogenetic analyses of dataset D1 under the three models that include only first and second codon position bases yield similar trees (scorpions as sister taxon to all other chelicerates; Hypothesis 4). Support for scorpions as sister taxon to all other

chelicerates decreases under the more realistic, partitioned *NTEp* model, suggesting that this placement may be artefactual. In contrast, the three models that include third codon position bases give trees with radically different placements of Scorpiones, and correspondingly different hypotheses regarding chelicerate evolution. In the *GTR+3p* analysis, inclusion of third codon position bases yields a clade of scorpions and *Argiope bruennichi*, a spider (Hypothesis 3). Hassanin (2006) has shown that *A. bruennichi* and *Euscorpius flavicaudis* both have an inverted pattern of mitochondrial strand-bias relative to other arthropods in third codon position bases. Analysis of the third codon position bases of *M. gibbosus* shows a similar pattern of skew, suggesting that the (Scorpiones + *A. bruennichi*) clade is an artefact of mitochondrial strand-bias. This proposition is supported by the results given by the *NTE+3* analysis, in which the portion of the third codon position signal due to mitochondrial strand-bias is excluded and scorpions are supported as sister taxon to the other chelicerates (Hypothesis 4). A final striking result is the change in the tree when the partitioned model is used (*NTE+3p*). Here, Xiphosura is identified as sister taxon to other chelicerates and Arachnida is monophyletic, with scorpions the sister taxon to other arachnids (Hypothesis 1/2). To test the significance of support for the favoured hypothesis, we can use the harmonic mean log likelihood of each hypothesis, calculated from the MrBayes runs, to estimate Bayes Factors as described in Kass and Raftery (1995). The Bayes Factors for [Hypothesis 1 vs Hypothesis 2] and [Hypothesis 1 vs Hypothesis 3] were 565.48 and 548.34 respectively, indicating very strong support for Hypothesis 1 (Bayes Factor values of >20 are considered to indicate strong support). In the same way, I calculated Bayes Factor support for a partitioned over a non-partitioned model

for the *NTE* and *NTE+3* analyses. The Bayes Factors were 1383.9 and 2670.62 respectively, indicating extremely strong support for the partitioned model in both pairs of analyses. I explain these results by suggesting that two potential sources of bias are present in the D1 dataset. Firstly, third codon position bases are strongly affected by mitochondrial strand-bias, which must be corrected for by using the NTE recoding scheme. Secondly, evolutionary model parameters vary between genes, leading to erroneous results if an underparameterised, unpartitioned model is used. Additionally, the commonly observed inversions of mitochondrial regions will lead inverted genes to acquire a strand bias opposite to that of genes in non-rearranged genomes. It is notable that robust support for the phylogenetic position of scorpions was only recovered from mitochondrial DNA sequences under the appropriate model. These findings emphasize the importance of model choice in phylogenetic analysis, and specifically the importance, when using multiple genes, of removing any sources of systematic bias and adequately allowing for differences in evolutionary models between genes.

3.5.2 D1 dataset skew

The CG skew results clearly show the varying evolutionary patterns of different mitochondrial genes in different taxa. Within a single species, CG skew direction varies between genes in a manner consistent with strand-bias effects and CG skew magnitude varies between codon positions in a manner consistent with greater redundancy at third positions. When examining multiple species, the same gene can have different directions of skew. This may be due to an inversion of the gene itself (in which case the other genes will have congruent skew across taxa, see Figure 3.4) or an inversion of the mitochondrial control region (in which case the other genes will also show opposite skew across taxa, see Figure 3.5). The widely differing patterns of evolution shown in Figure 3.9 graphically illustrate both the importance of partitioned models in analysis of mitochondrial genes, and the potential for strand-bias to mislead phylogenetic analysis even under well-parameterised models. The comparison of third codon position basis before and after NTE recoding demonstrates the effectiveness of the NTE coding scheme in reducing strand-bias, and supports the hypothesis that observed strand-bias mainly results from neutral and nearly neutral transitions.

The patterns of AT skew shown in Figure 3.10 do not seem to be attributable to strand-bias effects, since there is no correspondence between the direction of skew and the strand on which the gene resides. Rather, skew, where present, seems roughly constant across genes and taxa. This suggests that AT skew is driven by a coding strand effect (likely to be driven by codon usage patterns) rather than a mitochondrial genome strand effect, despite the strand-bias predicted by the replication model (Section 3.2.3).

These results underline the fact that the base composition of any given nucleotide sequence is likely to be the result of multiple different evolutionary pressures.

3.5.3 D2 dataset

Analysis of the D2 dataset under the *GTR+3p* and *NTE+3p* schemes showed the same pattern as observed for the D1 dataset (Table 3.2): *GTR+3p* gave Hypothesis 3 while *NTE+3p* gave Hypothesis 1/2. The poor resolution of the phylogeny recovered under the most realistic scheme, *NTE+3p* (Fig. 3.7) can probably be attributed to the large proportion of missing data, coupled with a loss of phylogenetic signal due to the NTE recoding scheme. While NTE removes erroneous phylogenetic signal due to strand-bias, it will also inevitably remove genuine phylogenetic signal.

A summary cladogram of the results of the phylogenetic analysis is given in Figure 3.11 showing the posterior probability assigned to each branch by D1 and D2 under the *NTE+3p* model and by W1. It is important to note that some orders, in particular other members of the clade Dromopoda (Opiliones, Scorpiones, Solifugae) as suggested by Schultz (1990) are not represented due to lack of sequence data. The addition of these taxa to the tree would permit more precise phylogenetic conclusions to be drawn, and distinguish between alternative Hypotheses 1 and 2. With that caveat, the summary cladogram, including orders represented in the D1 and D2 analyses, supports a traditional view of chelicerate evolution with Xiphosura a sister taxon to other

chelicerates and Arachnida monophyletic. Within the Arachnida, Scorpiones appear as a sister taxon to the remaining arachnids which form two high-level clades: ticks plus mites (Acari) and spiders plus whip spiders plus whip scorpions (Araneae + Uropygi + Amblypygi). Morphological, genomic, and developmental characters that are shared between scorpions and other arachnids presumably represent the ancestral arachnid state, allowing us to infer loss of such characters on the chelicerate phylogeny.

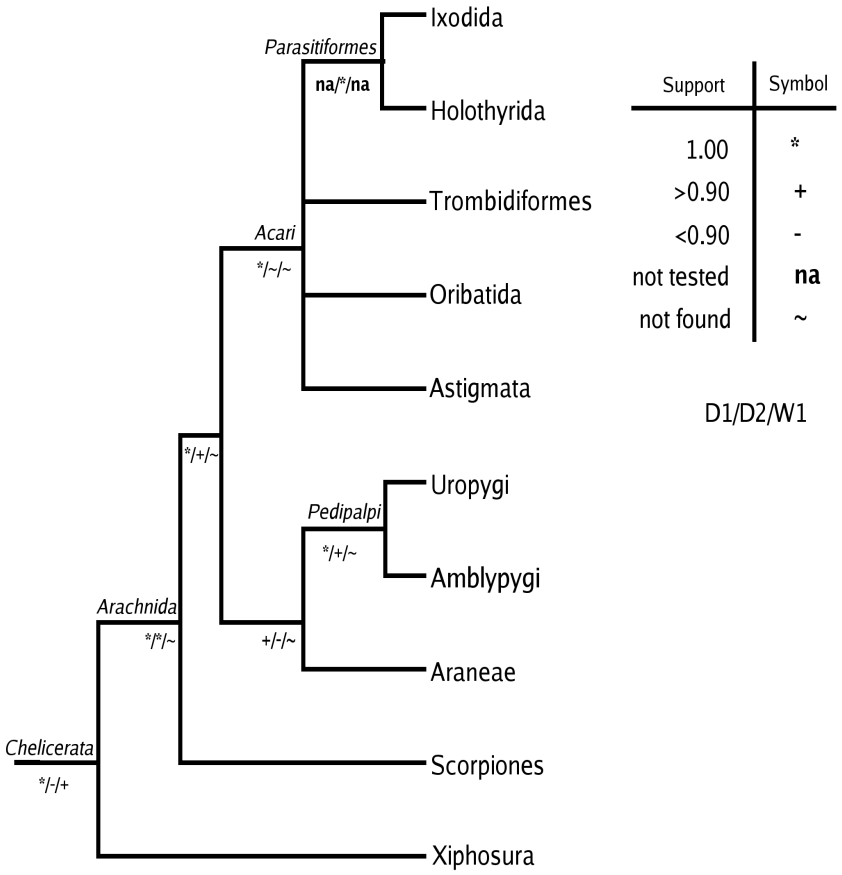


Figure 3.11: Summary cladogram of chelicerate relationships

Rooted summary cladogram of chelicerate orders assembled from analyses of the D1,D2 and W1 datasets. Pycnogonida is excluded due to its untenable placement in both datasets. Internal branches are named where they correspond to named taxonomic groups. Posterior probability support is shown below each branch, according to the key, for datasets D1/D2/W1. Support is given for a branch even if some taxa are missing from the dataset; for example, D1 does not include any representatives from Oribatida or Astigmata, but does support the grouping of Ixodida and Trombidiformes, therefore support has been shown for the branch labelled Acari.

3.5.4 The nuclear dataset

An interesting result is the apparent inability of nuclear rRNA alignments to robustly resolve interordinal relationships among the Chelicerata and plausible relationships between the various outgroup taxa, while robustly supporting relationships within orders (Figure 3.8). Nuclear rRNA has been used successfully for phylogenetic inference in other groups (Mallatt and Winchell 2002; Medina *et al.* 2001). One explanation for this is rapid cladogenesis at the ordinal level suggested by the D1 phylogeny (Figure 3.6 f) relative to the much longer period of stability following it, allowing only short periods for mutations to be fixed between speciation events. This scenario demonstrates the importance of using multiple genes (to recover more phylogenetic signal) and wide taxonomic sampling (to avoid long unbroken branches) when reconstructing troublesome phylogenies.

3.5.5 Missing data

As has been shown in many previous analyses, large datasets comprising multiple genes can be effective at resolving phylogenies in situations where single genes are insufficient, provided care is taken to avoid systematic bias (Philippe, Lartillot and Brinkmann 2005; Rokas *et al.* 2003). For multigene mitochondrial data, particular attention must be paid to normal and reverse strand-bias and to parameter differences between genes. Taxa can also be robustly placed using partial datasets. In D1, for

example, *A. bruennichi* is robustly placed despite having only 33% of characters present, and in D2, *E. flavicaudis* is robustly placed using only 38% of characters. It is likely that, in both of these cases, the confidence with which the taxa are placed can be attributed to two factors; a large absolute number of characters (4,857 coding nucleotides in the case of *A. bruennichi)* and the presence of a closely related species with a near-complete alignment in the dataset. In general, species not fulfilling these criteria are unlikely to be robustly placed by phylogenetic analysis of partial datasets. The sole pycnogonid species, *Endeis spinosa*, had ~40% of characters present in the D1 dataset, but had no closely related species in the alignment and was consistently placed in a clade of mites. More sequence data or a wider taxonomic sampling from the pycnogonids would be necessary to investigate this surprising result.

3.5.6 Implications for mitochondrial phylogenetics

Given the large historical interest in the use of mitochondrial genes as phylogenetic markers, particularly in the arthropods (Cameron, Barker and Whiting 2006; Cook, Yue and Akam 2005; Nardi *et al.* 2003), and the tendency towards the use of large, multiple gene datasets for phylogenetic reconstruction, the effect of model choice is a pertinent issue. The findings presented here indicate that, particularly in the case of deep phylogeny, mitochondrial gene sequences can be actively misleading. Extreme care should be taken, when using any multiple gene dataset, to avoid explore the potential for systematic bias. In the case of mitochondrial genes, strand bias should be of particular concern and the previous use of mitochondrial genomes in resolving deep phylogenies requires critical re-evaluation.

4 Phylogenetic analysis of Lophotrochozoa using nuclear and mitochondrial genes

4.1 Abstract

Of the three metazoan groups proposed by the new animal phylogeny – Lophotrochozoa, Ecdysozoa and Deuterostomia – Lophotrochozoa is the least phylogeneticaly resolved. However, many hypotheses regarding relationships between lophotrochozoan phyla have been advanced on the basis of morphology or small numbers of molecular characters. Here, I describe the assembly of a large set of aligned genes to address questions of relationship. For the three best-studied phyla, molluscs, annelids and platyhelminths, I examine the utility of nuclear protein coding, nuclear ribosomal RNA, mitochondrial protein coding and mitochondrial ribosomal RNA genes to resolve relationships. I also carry out combined analyses using all genes and investigate the contribution of heterotachy to the phylogenetic signal. I then look at the effect of adding representatives of less well-studied taxa, including some with very low numbers of aligned characters ('neglected taxa'). I find that analysis of individual and combined genes from the well-studied phyla largely fail to robustly resolve relationships, and that a strong signal of heterotachy is present in the combined analysis. Including additional taxa in the analysis greatly increases the resolution of the tree and supports (1) grouping of molluscs and annelids to the exclusion of platyhelminths, (2) close relationships between annelids, pogonophorans, echiurans and sipunculans, (3) grouping of rotifers and acanthocephalans as Syndermata, (4) a

sister taxon relationship between Cycliophora and Entoprocta and (5) grouping of Syndermata and Platyhelminthes as Platyzoa. Acoel flatworms do not group with Platyhelminthes but with Bryozoa. Other hypotheses, including Trochozoa and Lophophorata, are not recovered. These results show that taxa with only a small amount of sequence data can be robustly placed using a multigene approach, and that adequate taxon sampling is, in this case, more important than sequence completeness. However, computational limitations restrict the types of evolutionary models that can be applied to large numbers of taxa, which may limit ability to infer correct relationships.

4.2 Introduction

4.2.1 The origins of Lophotrochozoa

One of the most important hypotheses to arise from deep phylogenetic analysis of molecular data is the division of bilaterian animals into three superphyla – the Lophotrochozoa / Ecdysozoa / Deuterostomia (LED) hypothesis (Jones and Blaxter 2005). Suggested on the basis of single-gene studies (Adoutte *et al.* 1999; Aguinaldo *et al.* 1997; Halanych *et al.* 1995; Winnepenninckx *et al.* 1995) the LED hypothesis proposes a deep division between protostomes, in which the mouth develops from the primary blastopore, and deuterostomes, in which the mouth is derived from a secondary opening. Deuterostome phyla include chordates and ambulacarians (comprising the phyla Echinodermata, Hemichordata and Xenoturbellida). The protostomes are divided into two superphyla – Ecdysozoa (moulting animals),

including arthropods and nematodes, and Lophotrochozoa. Lophotrochozoa originally contained animals with a lophophore feeding structures (brachiopods, bryozoans and phoronids) and animals with trochozoan larvae (molluscs, annelids, nemerteans, sipunculans and echiurans) but has since been expanded to include platyhelminths (Adoutte *et al.* 1999; Balavoine 1997; Carranza, Baguna and Riutort 1997). The competing Acoelomata / Pseudocoelomata / Coelomata (APC) hypothesis proposes three superphyla based on the nature of the body cavity. Animals with a true body cavity lined with mesodermal tissue, including vertebrates, arthropods, molluscs and annelids, form Coelomata; those with a non-mesoderm-lined body cavity, including nematodes, form Pseudocoelomata; and those lacking a body cavity, including platyhelminths, form Acoelomata.

The LED hypothesis was initially supported by single-gene molecular studies and by developmental characters, but most multigene studies supported the APC hypothesis (Blair *et al.* 2002; Dopazo and Dopazo 2005; Wolf, Rogozin and Koonin 2004). However, a recent multigene study has taken steps to avoid systematic bias and found strong evidence in favour of LED (Philippe, Lartillot and Brinkmann 2005), though the positions of many phyla remain unresolved.

The relationships between phyla within these newly erected protostome clades have been a popular area for investigation with molecular data (Mallatt and Giribet 2006; Mallatt, Garey and Shultz 2004; Morris *et al.* 1996; Passamaneck and Halanych 2006; Petrov and Vladychenskaia 2005). However, molecular sequencing effort is far from equally divided among the two groups.

Taxon		nucleotide records	protein records	nuclear genome projects	Mitochondrial genome projects
Ecdysozoa		**5,465,885**	**375,251**	**5 (+3)**	**137**
	Arthropoda	3,950,107	292,577	4 (+1)	120
	Nematoda	1,508,059	82,385	1 (+2)	17
	Nematodmorpha	24	2	0	0
	tardigrada	6,923	55	0	0
	onycophora	236	185	0	0
	priapulida	536	47	0	0
Lophotrochozoa		**791,302**	**42,174**	**0**	**53**
	Mollusca	326,910	21,606	0	31
	Annelida / Echiura / Pogonophora	27,560	4,028	0	5
	Platyhelminthes	434,155	15,242	0	13
	Bryozoa	652	225	0	1
	Nemertea	400	179	0	0
	Brachiopoda	206	192	0	3
	Sipuncula	155	127	0	0
	Rotifera	1,264	575	0	0

Table 4.1: Comparison of sequencing effort in the principal phyla of Ecdysozoa and Lophotrochozoa

Numbers of nucleotide records, protein records and genome projects present in GenBank on 18/10/2006 are shown for each phylum, along with the total for Ecdysozoa and Lophotrochozoa. Numbers in parentheses represent genome projects not listed in the GenBank Genome database, but which are available at Ensembl. Not all phyla are shown.

Table 4.1 shows the numbers of nucleotide and protein records and the number of genome projects (mostly mitochondrial) for key ecdysozoan and lophotrochozoan phyla. There are ~4 times more nucleotide records, ~9 times more protein records and ~3 times more genome projects for Ecdysozoa than for Lophotrochozoa, in addition to whole nuclear genome projects. The majority of the disparity is due to the phylum Arthropoda, for which a vast amount of sequence data has been generated. In particular, the much greater number of genome projects for arthropod taxa mean that they have been the subject of several multigene studies (Cook, Yue and Akam 2005; Giribet *et al.* 2005; Hassanin 2006; Regier, Shultz and Kambic 2005). In contrast, multigene studies involving lophotrochozoans have generally been oriented toward testing support for the LED hypothesis rather than resolving lophotrochozoan

relationships (Blair *et al.* 2002; Philippe, Lartillot and Brinkmann 2005; Wolf, Rogozin and Koonin 2004). Of the three clades specified by the LED hypothesis, Lophotrochozoa is the least phylogenetically resolved.

4.2.2 Questions of Lophotrochozoan relationships

In a large group such as the Lophotrochozoa there are many possible relationships that could potentially be investigated.

Relationships between molluscs, platyhelminths and annelids

The three most well-studied lophotrochozoan phyla are the molluscs, platyhelminths and annelids, each of which contain species of medical or economic importance. Molluscs include bivalve species that are important in aquaculture as well as the large cephalopods (including squid and octopus). Annelids include parasitic leeches (class: Hirudinida) and terrestrial oligochaete worms, which play a key role in soil ecosystems. Platyhelminthes contains the parasitic tapeworms (class: Cestoda) as well as trematode parasites of humans (e.g. *Schistosoma spp.*). These three phyla are also of tremendous biological and evolutionary interest. Mollusca contains classes with incredibly diverse morphology and is an example of how modifications to a phylum body plan can generate diverse forms. Annelids are segmented, a feature that was previously used to group them with arthropods as Articulata. Their recently discovered lophotrochozoan affinity suggests that segmentation has arisen independently in

annelids and arthropods. Two minor phyla, spoonworms (Echiura) and tubeworms (Pogonophora) are often included in the annelids (Passamaneck and Halanych 2006). Platyhelminthes includes both parasitic and free-living species and phylogenetic studies of this group have provided insights about the evolution of parasitism (Littlewood, Rohde and Clough 1999; Olson and Tkach 2005). Platyhelminthes traditionally includes a group of very small flatworms lacking a gut (Acoelomorpha) but the acoels may not in fact be members of Platyhelminthes.

Traditional systematics based on morphology groups molluscs and annelids together as Trochozoa as they share trochophore-like larvae (Nielsen 1995). Platyhelminthes has been placed in a clade, Platyzoa, which also includes Rotifera, Acanthocephala and Gastrotricha. Platyzoa has traditionally been viewed as a basal bilaterian lineage (*vide* the APC hypothesis); however, molecular evidence has placed these taxa within the Lophotrochozoa (Adoutte *et al.* 1999; Balavoine 1997; Carranza, Baguna and Riutort 1997). Two hypotheses regarding the placement of Platyhelminthes have been advanced. Combined morphological and molecular (18S rDNA) evidence places Platyzoa as a sister group to Trochozoa (Giribet *et al.* 2000). A recent study using Bayesian inference to analyse combined 18S and 28S rDNA for lophotrochozoan taxa grouped platyhelminths and annelids to the exclusion of molluscs, albeit with marginal support (Passamaneck and Halanych 2006). Because platyhelminths, annelids and molluscs are all extremely well-represented in GenBank, we can test the relative support for Trochozoa versus (Annelida + Platyhelminthes) using multiple nuclear and mitochondrial genes and with relatively little missing data. **Acoelomorpha** is classified

within Platyhelminthes in the NCBI taxonomy, and this placement is supported by analysis of EF1A sequences (Berney, Pawlowski and Zaninetti 2000). However, a growing body of evidence (Giribet 2003; Littlewood, Rohde and Clough 1999; Ruiz-Trillo *et al.* 2004) suggests that acoel flatworms occupy a basal position in the bilaterian phylogeny. A relatively large number of nucleotide records for Acoelomorpha (983) mean that their placement can be tested with a greater number of characters than have previously been used.

The phylogenetic position of neglected phyla

Compared to the three well-studied phyla mentioned above, much smaller amounts of sequence data are available from members of other phyla. Nucleotide records for Rotifera (1,264), Bryozoa (652), Nemertea (400), Acanthocephala (346), Pogonophora (327), Cycliophora (253), Brachiopoda (206), Sipuncula (155) and Echiura (87), while small in number, represent a source of phylogenetic information and a test of the utility of the multigene approach. By including representatives of these phyla in an alignment a number of other phylogenetic hypotheses can be tested.

A monophyletic group, **Syndermata**, containing Acanthocephala and Rotifera has been proposed on the basis of molecular evidence. Acanthocephala have been placed either as a taxon within rotifers (Garey *et al.* 1996; Herlyn *et al.* 2003) using 18S rDNA sequences, or as sister taxon to rotifers (Passamaneck and Halanych 2006) in an analysis of combined 18S and 28S rDNA sequences. These hypotheses can be tested using a dataset that includes rotifer and acanthocephalan representatives. The inclusion of such taxa along with platyhelminths also allows us to test for monophyly

of **Platyzoa** (=Platyhelminthes, Acanthocephala and Rotifera).

The **Lophophorata** hypothesis unites Bryozoa and Brachiopoda (and Phoronida, unrepresented in this study) due to their possession of a lophophore, a feeding appendage with hollow tentacles. This morphological hypothesis has not been backed up by molecular data. Phylogenetic analysis of a nuclear-encoded gene (sodium–potassium ATPase) found Lophophorata paraphyletic (Anderson, Cordoba and Thollesson 2004), as did analysis of combined 18S and 28S rDNA (Passamaneck and Halanych 2006). This hypothesis can be tested using an alignment that includes bryozoan and brachiopod representatives.

Two neglected lophotrochozoan taxa, Nemertea and Sipuncula, are commonly grouped with molluscs and annelids under **Trochozoa** (Peterson and Eernisse 2001). We can test for the monophyly of Trochozoa (=Mollusca, Annelida, Nemertea and Sipuncula) using an alignment that includes representatives of all four phyla.

Annelida, Pogonophora and Echiura have been proposed to form a monophyletic group within the Trochozoa (Giribet *et al.* 2000) with Giribet *et al.* 2000 finding Annelida and Echiura as sister taxa with Pogonophora arising from the base of the group. Some workers (Halanych 2004; Passamaneck and Halanych 2006) found Echiura and Pogonophora to be within a paraphyletic Annelida.

The enigmatic phylum **Cycliophora**, represented by its single genus *Symbion*, has been placed close to entoprocts on the basis of nuclear ribosonal RNA genes (Passamaneck and Halanych 2006). An analysis including morphology and SSU data placed it as a sister taxon to Syndermata (Giribet *et al.* 2000). Inclusion of a representative of this phylum in the dataset will allow this hypothesis to be tested with

additional data.

Previous work

Previous studies of molecular lophotrochozoan phylogeny are characterised by their reliance on a small number of genes (summarised in Table 4.2)

		Conclusions						
Citation	**Character sets**	**Lophophorata**	**Platyzoa**	**Cycliophora + Entoprocta**	**Syndermata**	**Trochozoa**	**Annelida + Pogonophora + Echiura**	**Platyhelminthes**
Garey et al. 1996	18S ribosomal RNA	-	-	-	monophyletic[1]	-	-	-
Giribet et al. 2000	18S ribosomal RNA, 276 morphological characters	paraphyletic	monophyletic[2]	paraphyletic[3]	monophyletic	monophyletic	monophyletic	monophletic
Peterson and Eernisse 2001	18S ribosomal RNA, 138 morphological characters	monophyletic	-	-	-	monophyletic	monophyletic	not monophyletic[5]
Herlyn et al. 2003	18S ribosomal RNA	-	-	-	monophyletic[4]	-	-	-
Anderson, Cordoba and Thollesson 2004	Na-K ATPase a	paraphyletic	-	-	-	-	-	-
Passamaneck and Halanych 2006	18S ribosomal RNA, LSU ribosomal RNA	paraphyletic	monophyletic	monophyletic	monophyletic	paraphyletic	monophyletic	-

1 – Acanthocephala sister taxon to Bdelliodea
2 – With gastrotrichs and gnathostomulida and cycliophora
3 – Cycliophora sister taxon to Syndermata
4 – Acanthocephala sister taxon to Seisonidea
5 – Acoelomorpha not included in Platyhelminthes

Table 4.2: Conclusions of previous molecular studies of lophotrochozoan relationships

Columns give the citation, characters used in the analyses, and conclusions regarding the status of each hypothesis. A hyphen (–) indicates that the study did not address this question, or that there was no support to evaluate it.

Garey *et al.* (1996) found evidence, using SSU ribosomal RNA, for a monophyletic group containing Rotifera and Acanthocephala. Giribet *et al.* (2000) used SSU ribosomal RNA and morphological data to investigate triploblastic phyla, including several lophotrochozoan phyla. Peterson and Eernisse (2001) also used SSU ribosomal RNA and morphological characters in an analysis of 40 metazoan groups, although the most reliable trees excluded some taxa and some hypotheses could not be

tested. Herlyn *et al.* (2003) revisited the question of relationships within Syndermata using SSU ribosomal RNA. Anderson, Cordoba and Thollesson (2004) used a nuclear protein coding gene, sodium-potassium ATPase alpha subunit, to analyse metazoan relationships; however, only a few lophotrochozoan taxa were included. The most comprehensive molecular phylogeny of Lophotrochozoa was generated by Passamaneck and Halanych (2006), who used combined SSU and LSU ribosomal RNA sequence data for a range of taxa, allowing a range of hypotheses to be explicitly tested.

The success of the multigene approach in resolving questions of deep metazoan phylogeny suggests that such an approach might be useful for resolving lophotrochozoan relations. The relationships between phyla are the kind most likely to require the use of multiple genes to resolve since the cladogenesis events that must be reconstructed occurred very deep in evolutionary time and the branches between representatives of different phyla will necessarily be long. The numbers presented in Table 4.1 show that molecular sequence data is widely available for Lophotrochozoa and has been previously under-utilised in phylogenetics. Four lophotrochozoan phyla have >1,000 nucleotide records available, making it likely that a complete multigene dataset can be assembled for those phyla. In addition, less well-represented phyla may be included in a multigene analysis with a proportion of missing data present. In this chapter I describe the use of multiple genes to address questions of lophotrochozoan relationships. To test the relationships between Mollusca, Annelida and Platyhelminthes I assembled a dataset consisting of mitochondrial protein-coding, nuclear protein-coding, mitochondrial ribosomal RNA and nuclear ribosomal RNA

genes for a selection of taxa with high sequence representation in GenBank (L1). To further test lophotrochozoan relationships, including those concerning neglected phyla, I assembled a further dataset consisting of mitochondrial protein-coding, mitochondrial ribosomal RNA and nuclear ribosomal RNA genes for a wide selection of lophotrochozoan taxa (L2). Nuclear protein-coding genes were not included in this dataset owing to their much lower representation in GenBank.

4.3 Methods

4.3.1 Data collection

The TaxMan software package, described in detail in Chapter 2, was used to assemble a dataset of aligned protein and DNA sequences for lophotrochozoan taxa. For a detailed discussion of sequence extraction, consensus building and alignment of this dataset, see Section 2.6.2. Mitochondrial protein coding and ribosomal RNA genes, and nuclear ribosomal RNA genes, were included in the gene set *a priori* since they are known to be well-represented in GenBank. To identify well-represented nuclear genes, TaxMan was used to search for common gene names (see Section 2.5.1), identifying ACTIN, H3 and EF1A as well-represented genes. For ACTIN, orthology is less certain than for the mitochondrial protein-coding genes, since multiple copies may be present in a genome (for example, *Schistosoma mansoni* has been shown to have two ACTIN genes [Oliveira and Kemp 1995]). However, all three nuclear genes have been used in previous phylogenetic studies (e.g. Berney, Pawlowski and Zaninetti 2000; Cadez, Raspor and Smith 2006). Since an aim of this study was to compare the

resolution afforded by genes from different organelles, all three genes were included in the dataset.

4.3.2 Phylogenetic analysis

Choosing datasets

Sets of taxa were chosen for analysis using the slice feature in TaxMan (Section 2.3.6). For the set of genes included in the dataset (see Section 2.6.2), TaxMan calculated sequence completeness for each species and chose the best representatives of each class. These choices were manually refined to ensure good taxonomic sampling and a computationally tractable number of species. For the L2 dataset, only species with at least 4,000 aligned characters were included. Because many more genes were included than in previous studies, and because taxon selection was made automatically on the basis of sequence completeness, the taxa included were different from those used in previous studies.

Evolutionary Models

For each gene, MrModeltest (Nylander 2004), a modified version of Modeltest (Posada and Crandall 1998) was used to evaluate the fit of various models implemented by

Gene	Model selected by AIC
ACTIN	GTR+G
H3	GTR+I
EF1A	GTR+I+G
ATP6	GTR+I+G
COX1	GTR+I+G
COX2	GTR+I+G
COX3	GTR+I+G
CYTB	GTR+I+G
ND1	GTR+I+G
ND2	GTR+I+G
ND3	GTR+I+G
ND4L	GTR+I+G
ND4	GTR+I+G
ND5	GTR+I+G
ND6	GTR+I+G
RNA_12S	GTR+I+G
RNA_16S	GTR+G
RNA_LSU	GTR+I+G
RNA_SSU	GTR+G

Table 4.3: AIC selection of evolutionary models for genes in the L1 dataset

MrBayes using the alignment from the set of species in L1 (Table 4.3). ATP8 was excluded from model testing and from subsequent phylogenetic analysis as it is very short, and alignment was problematic. For every gene the model selected by the Akaike Information Criterion (AIC; Posada and Buckley 2004) was the General Time Reversible model (GTR) with either gamma rate variation (GTR+G), a proportion of invariant sites (GTR+I) or both (GTR+I+G). Since Bayesian inference is relatively robust to overparameterisation (Lemmon and Moriarty 2004), for phylogenetic analysis a GTR+G+I model was applied to all genes. Alignments were partitioned by

gene, with base frequencies, substitution rates, alpha parameters and proportions of invariant sites unlinked across partitions, and a rate multiplier used to allow rate variation between partitions. Model details specific to individual analyses are given in the Results.

Selection of conserved blocks

Gblocks 0.9b (Castresana 2000) was used to select conserved blocks from the alignments for analysis using the following parameters:

for the L1 dataset:

Minimum Number Of Sequences For A Conserved Position: **9**

Minimum Number Of Sequences For A Flanking Position: **9**

Maximum Number Of Contiguous Nonconserved Positions: **10**

Minimum Length Of A Block: **5**

Allowed Gap Positions: **All**

for the L2 dataset:

Minimum Number Of Sequences For A Conserved Position: **20**

Minimum Number Of Sequences For A Flanking Position: **20**

Maximum Number Of Contiguous Nonconserved Positions: **10**

Minimum Length Of A Block: **5**

Allowed Gap Positions: **All**

Tree reconstruction

Alignments were analysed using MrBayes 3.1 (Ronquist and Huelsenbeck 2003). Default MCMCMC parameters were used with the following exceptions: the number of chains per run was set to 8, and the Dirichlet tuning parameter for the rate multiplier

was increased from 500 to 50,000. Using eight chains per run rather than four leads to faster convergence of independent runs in preliminary analyses (data not shown). Increasing the Dirichlet tuning parameter for the rate multiplier lead to better exploration of parameter space (shown by a higher proportion of accepted changes in the rate multiplier parameter; ~40% rather than ~0.3%, data not shown). The parallel version of MrBayes was used to spread analysis over 16 nodes in a high performance compute cluster (3.4 GHz Intel Xeon processors with 500MB equivalent RAM per node). During parallel analysis, runs were periodically monitored using Tracer (web reference 12) to determine convergence. Once convergence of split frequencies, convergence of parameter estimates and flattening of likelihood scores were confirmed, the burn-in period was noted and the runs allowed to continue for an additional 500,000 generations. Trees sampled after the burn-in period were summarised to give the 50% majority rule consensus tree.

Evolutionary distances

To identify slowly-evolving genes, dnadist from the PHYLIP (Felsenstein 2005) package was used to calculate the average LogDet (Lockhart *et al.* 1994) distances between taxa for each gene.

4.4 Results

4.4.1 Data gathering

Table 4.4 shows details of the aligned dataset gathered by TaxMan. As expected for large taxonomic groups, the distribution of sequence data is highly skewed with respect to both genes and species. Molluscs, annelids and platyhelminths, which had the most raw sequence data available (Table 4.1) are also the best-represented phyla in terms of aligned consensus sequences. This characteristic pattern of skewed sequence distribution is seen both at the phylum level and at the class level (classes within Mollusca, for example). Notably, due to increased redundancy in large collections of sequences, the greater than ten-fold difference in the numbers of raw sequence records between annelids and molluscs translates into an approximately five-fold difference in the number of aligned consensus sequences.

The pattern of distribution across genes is slightly different: four genes commonly used in phylogenetics (COX1, RNA_16S, RNA_SSU and RNA_LSU) have roughly similar numbers of sequences while the remaining genes have many fewer sequences.

Phylum	Class	no. taxa	COX1	RNA_16S	RNA_SSU	RNA_LSU	RNA_12S	ND1	H3	CYTB	COX3	ACTIN	EF1A	COX2	ND4	ND4L	ND6	ATP6	ND3	ND2	ND5	ATP8	*All genes*
Mollusca		**5026**	**2913**	**2780**	**1097**	**1366**	**606**	**344**	**250**	**235**	**143**	**160**	**56**	**80**	**68**	**76**	**48**	**39**	**35**	**37**	**44**	**23**	***10400***
	Gastropoda	3443	2107	2030	622	869	437	210	103	169	19	104	11	31	12	55	27	15	10	12	20	11	*6874*
	Bivalvia	1144	536	477	340	341	46	122	58	23	52	11	14	25	44	10	9	12	13	14	12	1	*2160*
	Cephalopoda	363	217	238	79	113	122	9	53	40	69	44	30	21	9	8	9	9	9	8	9	9	*1105*
	Polyplacophora	38	30	27	30	30	1	1	28	1	1	1	1	1	1	1	1	1	1	1	1	0	*159*
	Scaphopoda	29	18	3	19	6	0	2	2	2	2	0	0	2	2	2	2	2	2	2	2	2	*72*
	Aplacophora	9	5	5	7	7	0	0	6	0	0	0	0	0	0	0	0	0	0	0	0	0	*30*
Platyhelminthes		**1661**	**277**	**126**	**1110**	**1014**	**50**	**100**	**23**	**22**	**19**	**7**	**46**	**19**	**20**	**13**	**20**	**25**	**26**	**22**	**16**	**0**	***2955***
	Trematoda	598	111	25	362	434	16	77	5	7	6	4	6	7	8	7	12	7	15	9	6	0	*1124*
	Cestoda	400	94	81	268	259	33	19	1	8	6	2	23	6	10	6	7	13	11	10	6	0	*863*
	Monogenea	360	27	11	228	205	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	*472*
	Turbellaria	303	45	9	252	116	1	4	17	7	7	1	16	6	2	0	1	5	0	3	4	0	*496*
Annelida		**880**	**382**	**289**	**546**	**293**	**126**	**79**	**42**	**22**	**13**	**2**	**32**	**22**	**6**	**6**	**11**	**7**	**7**	**7**	**6**	**20**	***1918***
	Polychaeta	435	126	144	312	139	4	3	40	17	7	0	19	3	3	3	7	3	3	3	3	3	*842*
	Clitellata	276	105	142	112	91	49	18	2	3	4	1	7	17	3	3	3	3	3	3	3	16	*588*
	Hirudinida	148	131	3	102	61	73	58	0	2	2	1	6	2	0	0	1	1	1	1	0	1	*446*
	Branchiobdellae	21	20	0	20	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	*42*
Bryozoa		**95**	**22**	**74**	**23**	**6**	**10**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	***136***
	Gymnolaemata	79	22	64	13	3	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	*103*
	Phylactolaemata	13	0	10	8	1	9	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	*29*
	Stenolaemata	3	0	0	2	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	*4*
Nemertea		**116**	**63**	**74**	**49**	**63**	**0**	**0**	**42**	**0**	**0**	**0**	**4**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	***295***
	Enopla	72	44	37	36	37	0	0	28	0	0	0	1	0	0	0	0	0	0	0	0	0	*183*
	Anopla	44	19	37	13	26	0	0	14	0	0	0	3	0	0	0	0	0	0	0	0	0	*112*
Rotifera		**92**	**71**	**23**	**65**	**58**	**0**	**0**	**32**	**1**	**1**	**1**	**1**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	***254***
	Bdelloidea	19	15	2	9	7	0	0	2	1	1	0	1	1	0	0	0	0	0	0	0	0	*39*
	Monogononta	71	55	21	54	50	0	0	30	0	0	1	0	0	0	0	0	0	0	0	0	0	*211*
	Seisonidea	2	1	0	2	1	0	0		0	0	0	0	0	0	0	0	0	0	0	0	0	*4*
Brachiopoda		**72**	**36**	**8**	**45**	**8**	**15**	**5**	**2**	**5**	**5**	**0**	**1**	**5**	**5**	**3**	**5**	**5**	**5**	**4**	**5**	**3**	***170***
	Rhynchonellata	51	30	7	27	3	14	3	0	3	3	0	0	3	3	1	3	3	3	2	3	2	*113*
	Lingulata	9	3	0	8	1	0	1	0	1	1	0	1	1	1	1	1	1	1	1	1	1	*25*
	Phoroniformea	8	2	1	7	3	1	1	2	1	1	0	0	1	1	1	1	1	1	1	1	0	*27*
	Craniata	3	1	0	2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	*4*
	unclassified Brachiopoda	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	*1*
Acanthocephala		**62**	**33**	**3**	**53**	**29**	**1**	**1**	**2**	**1**	**1**	**0**	**0**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**0**	***132***
	Archiacanthocephala	11	5	2	8	6	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	***21***
	Polyacanthocephala	1	1	0	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	***3***
	Palaeacanthocephala	44	25	1	38	20	1	1	2	1	1	0	0	1	1	1	1	1	1	1	1	0	***98***
	Eoacanthocephala	6	2	0	6	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	***10***
Sipuncula		**38**	**15**	**2**	**30**	**28**	**1**	**0**	**26**	**1**	**1**	**0**	**1**	**1**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**1**	***108***
	Phascolosomatidea	18	6	0	16	16	0	0	15	0	0	0	0	0	0	0	0	0	0	0	0	0	*53*
	Sipunculidea	17	9	2	13	11	1	0	10	1	1	0	1	1	0	0	1	0	0	0	0	1	*52*
	unclassified Sipuncula	3	0	0	1	1	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	*3*
Pogonophora		**36**	**23**	**13**	**20**	**8**	**1**	**2**	**1**	**2**	**2**	**0**	**3**	**2**	**1**	**0**	**2**	**1**	**2**	**2**	**0**	**2**	***87***
	Vestimentifera	19	17	5	7	7	1	1	1	1	1	0	3	1	1	0	1	1	1	1	0	1	*51*
	Perviata	14	5	8	11	1	0	1	0	1	1	0	0	1	0	0	1	0	1	1	0	1	*33*
	unclassified Pogonophora	3	1	0	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	*3*
Echiura		**9**	**2**	**2**	**5**	**3**	**0**	**1**	**2**	**1**	**1**	**0**	**2**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	***28***
Cycliophora		**5**	**3**	**1**	**2**	**3**	**0**	**0**	**2**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	***11***
Entoprocta		**5**	**1**	**0**	**5**	**2**	**0**	**0**	**1**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	**0**	***9***
	All classes	***8097***	***3841***	***3395***	***3050***	***2881***	***810***	***532***	***425***	***291***	***186***	***170***	***146***	***132***	***102***	***100***	***89***	***79***	***77***	***74***	***73***	***50***	***16503***

Table 4.4: Aligned gene sequences gathered by TaxMan for Lophotrochozoa (legend overleaf)

Table 4.4

Columns list the number of taxa, number of consensus sequences for each gene, and number of sequences for all genes for each class. Numbers in bold show the total for each phylum. Bars at the end of each column show the total number of sequences for that group. Column totals and bars show the total number of sequences for each gene.

4.4.2 L1 dataset – Molluscs, Annelids and Platyhelminths

Table 4.5 shows the species included in the L1 dataset.

NCBI taxid	Phylum	Class	Species	Characters
6359	Annelida	Polychaeta	*Platynereis dumerilii*	15005
147105	Annelida	Branchiobdellae	*Cambrincola pamelae*	3051
35632	Annelida	clitellata	*Lumbricus rubellus*	18639
279730	Annelida	Hirudinida	*Haementeria depressa*	6976
211838	Mollusca	Aplacophora	*Chaetoderma sp.*	4772
34587	Mollusca	Polyplacophora	*Katerina tunicata*	14607
6610	Mollusca	Cephalopoda	*Sepia officinalis*	16868
6500	Mollusca	Gastropoda	*Aplysia californica*	19998
6577	Mollusca	Bivalvia	*Plactopecten magellanicus*	20148
203167	Mollusca	Scaphopoda	*Siphonodentalium lobatum*	12251
90914	Platyhelminthes	Turbellaria	*Paratomella rubra*	12699
54594	Platyhelminthes	Monogenea	*Polystomoides malayi*	6510
6182	Platyhelminthes	Trematodda	*Schistosoma japonicum*	22517
6216	Platyhelminthes	Cestoda	*Hymenolepis diminuta*	18749
6669	Arthropoda	Branchiopoda	*Daphnia pulex*	19461

Table 4.5: Details of species included in the L1 dataset

Columns give the NCBI taxonomic id, the phylum and class to which the species belongs, the binomial name and the total number of nucleotide characters in all genes. Species were automatically selected by TaxMan such that each class is represented by the species with the most complete set of sequences.

The genes included in the L1 dataset were split into four groups:

Mitochondrial protein-coding genes: ATP6, COX1, COX2, COX3, CYTB, ND1, ND2, ND3, ND4, ND4L, ND5, ND6

Mitochondrial ribosomal RNA genes: RNA_12S, RNA_16S

Nuclear protein-coding genes: ACTIN, EF1A, H3

Nuclear ribosomal RNA genes: RNA_SSU, RNA_LSU

Genes from different organelles (nuclear vs. mitochondrial) may be subject to different types of bias and carry different phylogenetic signals. The same is true for protein-coding vs. ribosomal RNA genes. To investigate the utility of different types of genes for lophotrochozoan phylogenetics, each group of genes was analysed individually, before a combined analysis was carried out. In all Figures, outgroup species are labelled grey.

Single group analyses – nuclear protein-coding genes

The nuclear protein-coding genes were the least well-represented in the L1 dataset. Of the sixteen species in the L1 dataset (Table 4.5), only five (*S. japonicum, A. californica, L. rubellus* and the two outgroup species) had sequences available for all three genes. Two had sequences available for two genes (*S. officinalis* for ACTIN and H3; *P. dumerilii* for EF1A and H3). Four had sequences available for only a single gene (*P. magellanicus* for ACTIN; *H. depressa* and *H. diminuta* for EF1A; *K. tunicata* for H3). Five species (*C. panelae, Chaetoderma sp., S. lobatum, P. rubra,* and *P. malayi*) had no sequence data available for nuclear protein-coding genes and were excluded from the analysis. Because the sequences were derived from EST data, full-length translations could not be made and codon positions could not be easily identified, and because of the large amount of missing data, Gblocks was unable to identify any well-conserved positions. Therefore, all positions were used in the analysis.

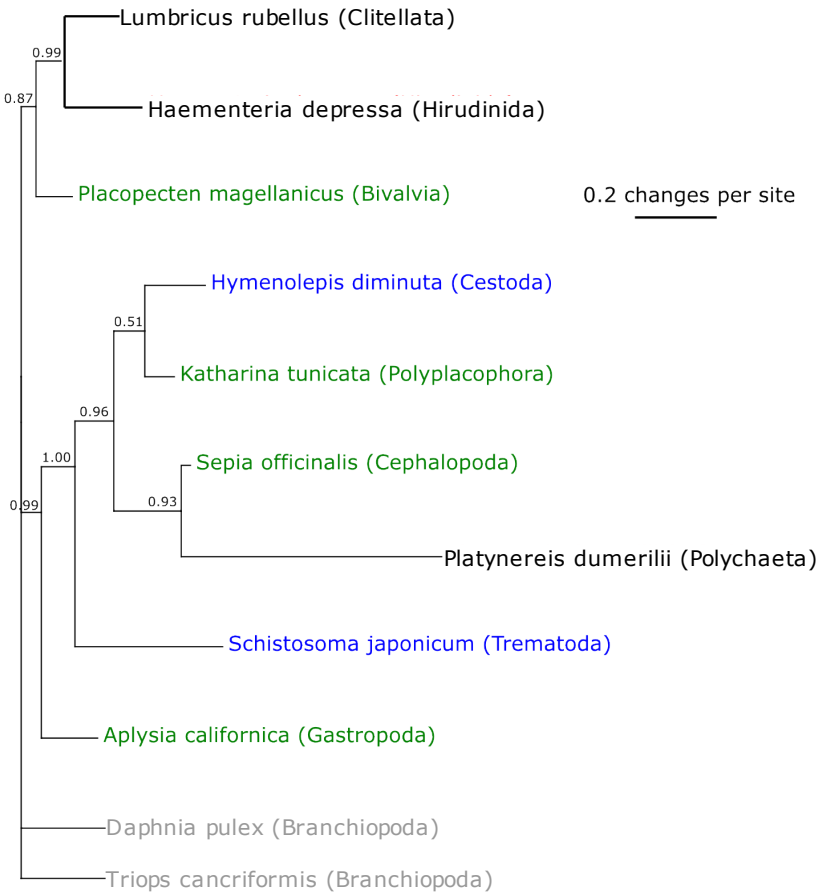


Figure 4.1: Tree derived from Bayesian analysis of nuclear protein-coding genes from the L1 dataset

Terminal nodes are labelled with the binomial name and the class in parentheses. Species are coloured according to phylum: black – Annelida, green – Mollusca, blue – Platyhelminthes. The scale bar shows the branch length associated with 0.2 expected changes per site.

The tree was well-resolved but failed to recover monophyly of any of the three phyla represented (Figure 4.1). Clitellata and Hirudinida were grouped with high posterior probability (99%). A number of unexpected clades were recovered with high support: (1) Cestoda + Polyplacophora + (Cephalopoda + Polychaeta); (2) 1 + Trematoda; (3) 2 + Gastropoda.

Single group analyses – nuclear ribosomal RNA genes

In contrast to the nuclear protein coding genes, nuclear ribosomal genes were well-represented in the dataset. All species in L1 had both genes represented, with the exception of *P. dumerilii* (class: Polychaeta, missing RNA_SSU and RNA_LSU), *H. depressa* (class: Hirudinida, missing RNA_SSU) and *S. lobatum* (class: Scaphopoda, missing RNA_LSU). These three species were excluded from the analysis. To ensure that the polychaete worms were represented, the species with the most sequence data for these genes, *Nereis pelagica,* was included as a representative of Polychaeta. Gblocks was used to select conserved positions for analysis (see methods, Section 4.3.2).

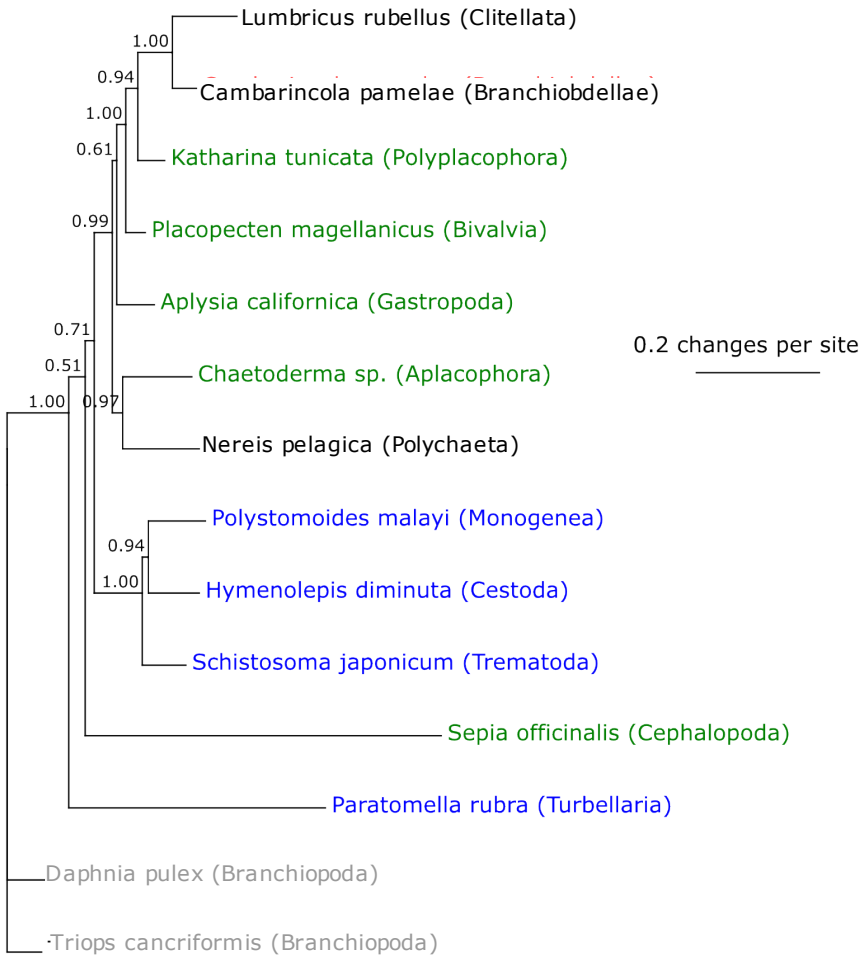


Figure 4.2: Tree derived from Bayesian analysis of nuclear rRNA genes from the L1 dataset

Terminal nodes are labelled with the binomial name and the class in parentheses. Species are coloured according to phylum: black – Annelida, green – Mollusca, blue – Platyhelminthes. The scale bar shows the branch length associated with 0.1 expected changes per site.

Analysis of nuclear ribosomal RNA genes yielded a fully resolved tree, although support for some branches was low (Figure 4.2). Annelida and Mollusca were mutually paraphyletic. Platyhelminthes was paraphyletic due to the placement of *P. rubra*, an acoel turbellarian, near the base of the tree. The branch supporting the

grouping of the other three platyhelminth orders, Monogenea, Cestoda and Trematoda, was strongly supported. The cephalopod *S. officinalis* was placed near the base of the tree.

The placement of *S. officinalis* and *P. rubra* near the base of the tree, although with only limited support, could be a long-branch attraction artefact. Reanalysis of the nuclear ribosomal RNA genes with these two taxa excluded yielded a tree with higher support values (Figure 4.3). In this tree there was 100% posterior probability support for a grouping of annelids and molluscs to the exclusion of platyhelminths, although Mollusca and Annelida remained paraphyletic. 100% posterior probability was found for monophyletic Platyhelminthes.

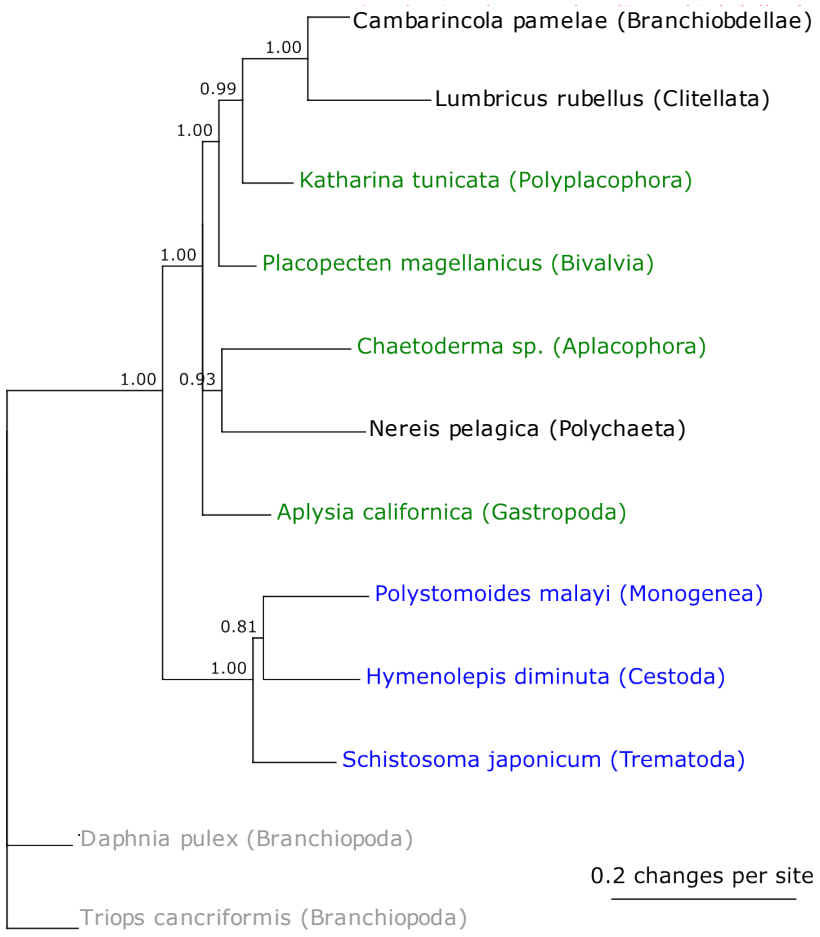


Figure 4.3: Tree derived from Bayesian analysis of nuclear ribosomal RNA genes with long-branch taxa excluded.

Terminal nodes are labelled with the binomial name and the class in parentheses. Species are coloured according to phylum: black – Annelida, green – Mollusca, blue – Platyhelminthes. The scale bar shows the branch length associated with 0.2 expected changes per site.

Single group analyses – mitochondrial ribosomal RNA genes

Mitochondrial ribosomal RNA genes were less-well represented than the nuclear ribosomal RNA genes in the L1 dataset. *C. pamelae*, *Chaetoderma sp.* and *P. malayi* were missing both genes and were excluded from the analysis. *P. dumerilii* was missing RNA_16S only. *H. depressa, S. lobatum* and *P. rubra* were missing

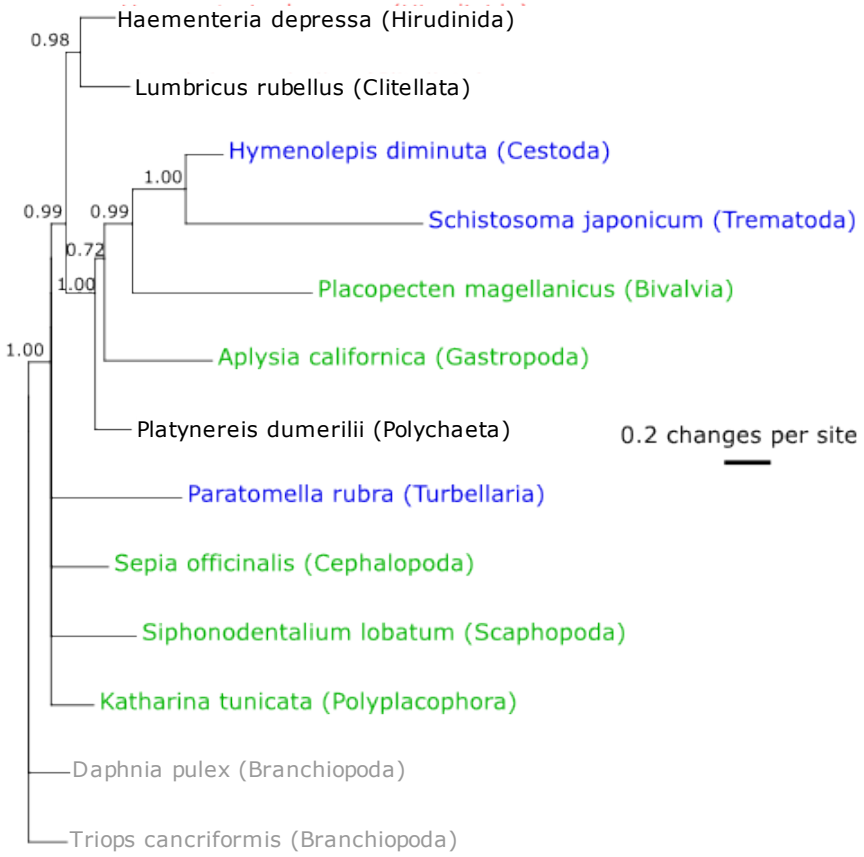


Figure 4.4: Tree derived from Bayesian analysis of mitochondrial rRNA genes from the L1 dataset

Terminal nodes are labelled with the binomial name and the class in parentheses. Species are coloured according to phylum: black – Annelida, green – Mollusca, blue – Platyhelminthes. The scale bar shows the branch length associated with 0.2 expected changes per site.

RNA_12S only. Gblocks was used to select conserved positions for inclusion in the analysis.

Analysis of mitochondrial ribosomal RNA genes yielded a poorly resolved tree (Figure 4.4). None of the three phyla were monophyletic. The tree proposed sister-taxon relationships between Hirudinida and Clitellata, and between Cestoda and Trematoda, both with high support. However, it also proposed an unlikely sister-taxon relationship between (Cestoda+Trematoda) and bivalve molluscs. Two other surprising clades were given high support: (1) a clade containing cestodes, trematodes, bivalves, gastropods and polychaetes and (2) a clade containing (1) plus Hirudinida and Clitellata to the exclusion of the three remaining mollusc classes (cephalopods, scaphopods and polyplacophorans).

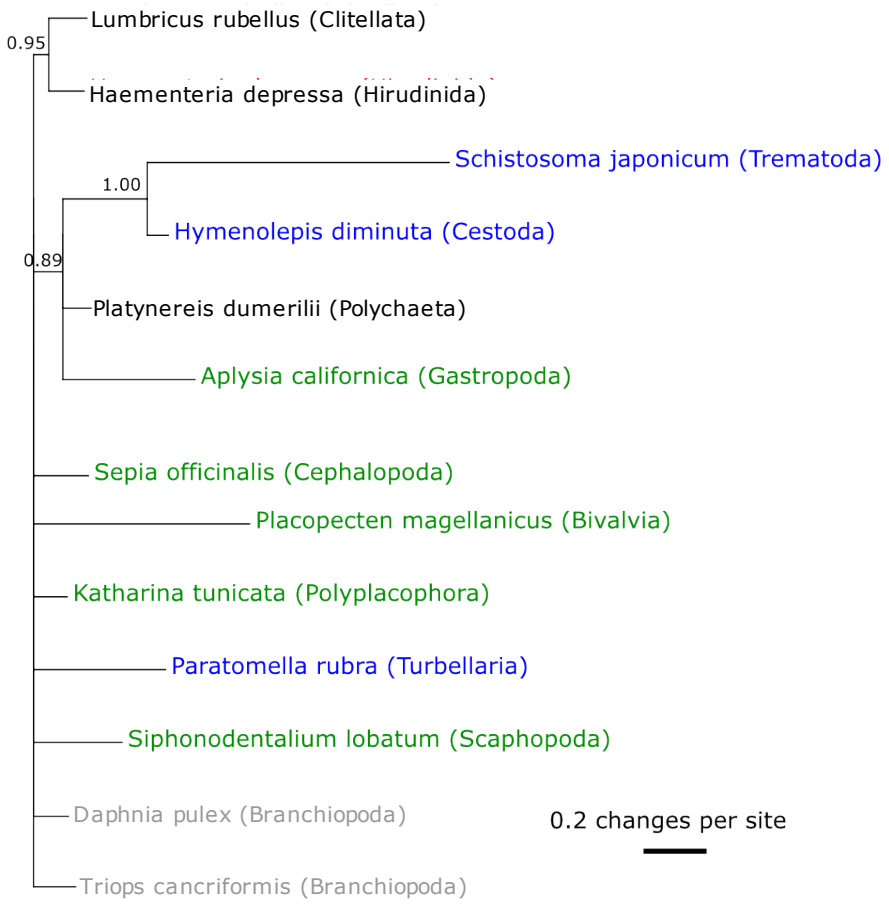


Figure 4.5: Tree derived from Bayesian analysis of R/Y-recoded mitochondrial ribosomal RNA genes

Terminal nodes are labelled with the binomial name and the class in parentheses. Species are coloured according to phylum: black – Annelida, green – Mollusca, blue – Platyhelminthes. The scale bar shows the branch length associated with 0.2 expected changes per site.

To investigate the effect of mitochondrial strand-bias on this analysis, the same dataset was analysed with bases recoded as purine or pyrimidine (R/Y). For this analysis, a Jukes-Cantor model (JC; nst=1) was used. In the resulting tree (Figure 4.5), support

for the unexpected relationships listed above was reduced or absent. Support for (Trematoda+Cestoda) was unchanged at 100%, while support for (Clitellata + Hirudinida) reduced from 98% to 95%.

Single group analyses – mitochondrial protein-coding genes

For mitochondrial genes, all but one species (*Chaetoderma sp.*, which was excluded from the analysis) in the L1 dataset had some sequence data. The numbers of genes present ranged from twelve (all present, *D. pulex*) to one (only COX1 available, *C. pamelae* and *P. malayi*). Analysis of mitochondrial protein-coding genes gave a partially-resolved tree in which annelids (but not molluscs or platyhelminths) were monophyletic (Figure 4.6).

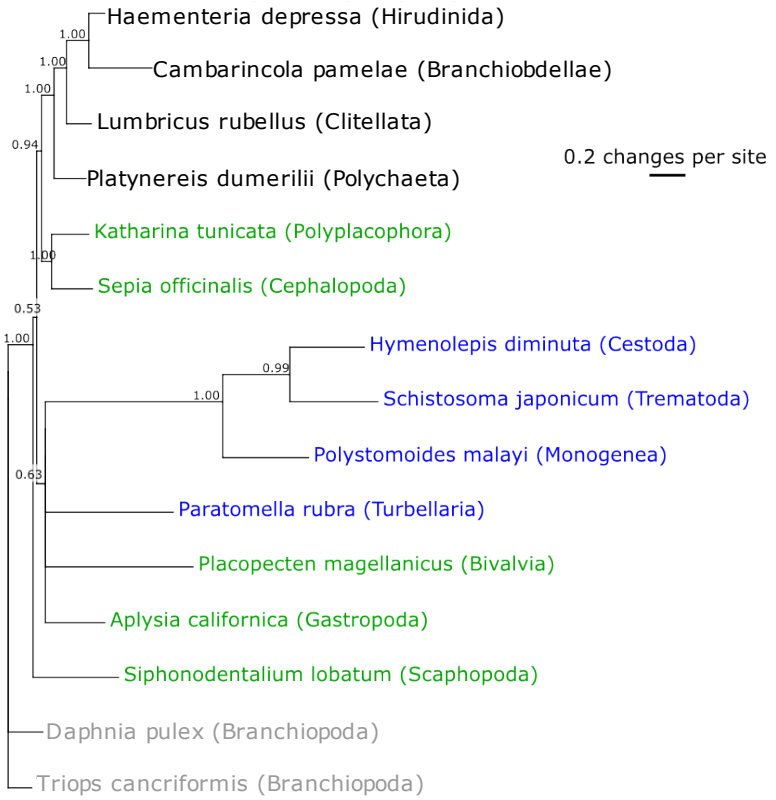


Figure 4.6: Tree derived from Bayesian analysis of mitochondrial protein-coding genes from the L1 dataset

Terminal nodes are labelled with the binomial name and the class in parentheses. Species are coloured according to phylum: black – Annelida, green – Mollusca, blue – Platyhelminthes. The scale bar shows the branch length associated with 0.2 expected changes per site.

Annelid monophyly was given 100% posterior probability, and resolution within the annelids was high, supporting (((Hirudinida+Branchiobdellae)+Clitellata)+Polychaeta). Strong support was also found for the grouping of cephalopods and polyplacophorans, and for grouping of the parasitic platyhelminths (cestodes, trematodes and monogeneans). A sister taxon

relationship was found between the annelids and the (Cephalopoda + Polyplacophora) clade. Because mitochondrial strand-bias has been shown to affect phylogenetic reconstruction, this analysis was repeated with the data matrix NTE-recoded according to Hassanin 2005 (Figure 4.7). This tree showed strong support for groups present in the non-recoded tree: annelids, parasitic platyhelminths, (cephalopods+polyplacophorans), and (cephalopods+polyplacophorans) + annelids. Relationships within the annelids and parasitic platyhelminths were unchanged. The NTE-recoded tree showed support for a clade of bivalves and free-living platyhelminths (Turbellaria). Notably, the branch leading to the parasitic platyhelminths, and branches relating them, were long relative to other taxa.

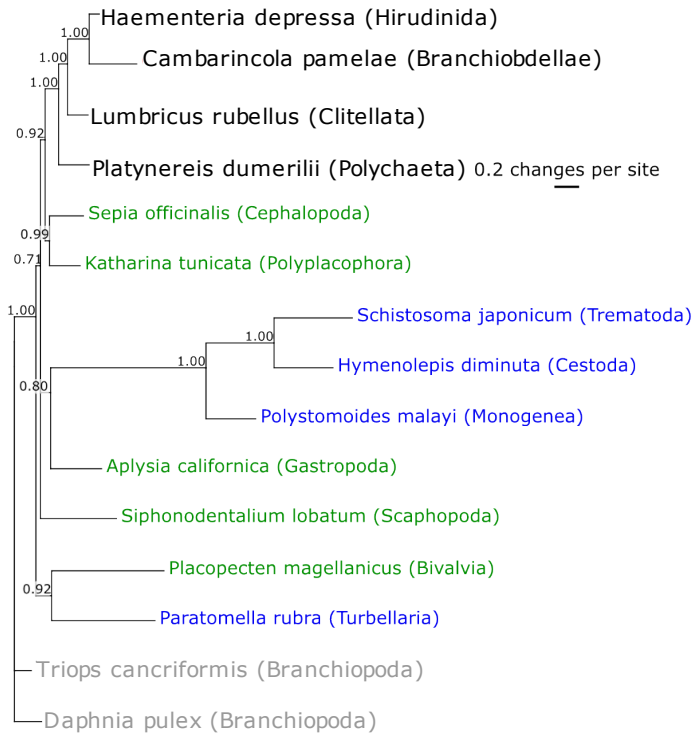


Figure 4.7: Tree derived from analysis of NTE-recoded mitochondrial protein-coding genes from the L1 dataset

Terminal nodes are labelled with the binomial name and the class in parentheses. Species are coloured according to phylum: black – Annelida, green – Mollusca, blue – Platyhelminthes. The scale bar shows the branch length associated with 0.2 expected changes per site.

Combined dataset

The combined dataset included all genes for the L1 taxa. Third codon position bases were excluded and only conserved positions, selected by Gblocks, were included for all genes except nuclear protein-coding genes. Because of the inability to robustly identify codon positions, and the inability of Gblocks to identify conserved residues, all positions were included for the nuclear protein-coding genes (ACTIN, H3, EF1A). To avoid mitochondrial strand-bias effects, mitochondrial protein-coding genes were recoded according to the NTE scheme, while mitochondrial RNA genes were recoded as purine/pyrimidine. The JC (nst=1) model was applied to the mitochondrial RNA genes while a GTR model was applied to other genes. Partitioning and other model parameters were as given in the methods (Section 4.3.2).

The tree was partially resolved (Figure 4.8). Strong support was found for monophyly of the parasitic platyhelminths (cestodes, trematodes and monogeneans) and, within that clade, for (cestodes+trematodes), although platyhelminths as a whole were paraphyletic, with Turbellaria sister taxon to the remaining ingroup species. Hirudinida, Clitellata and Branchiobdellae were grouped with strong support, although the remaining annelid class, Polychaeta, was sister taxon to Cephalopoda, making Annelida paraphyletic. Other relationships were not well-supported.

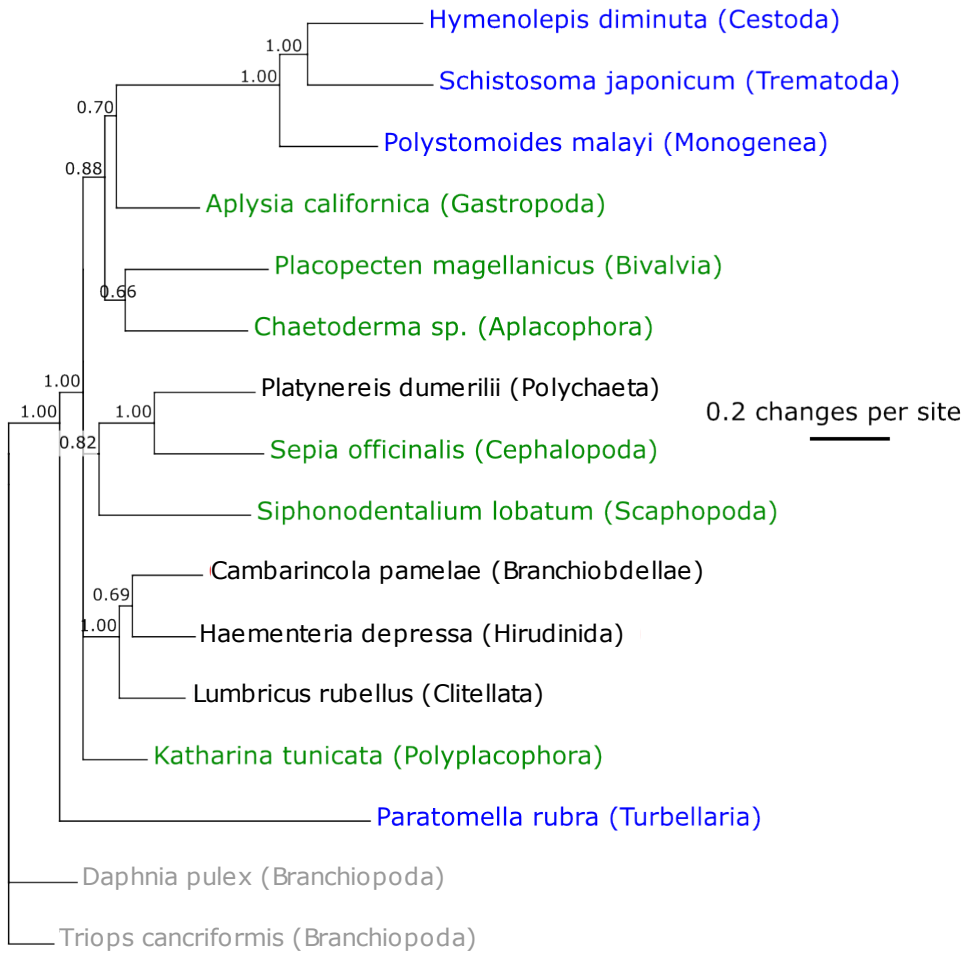


Figure 4.8: Tree derived from Bayesian analysis of all genes for the L1 dataset

Terminal nodes are labelled with the binomial name and the class in parentheses. Species are coloured according to phylum: black – Annelida, green – Mollusca, blue – Platyhelminthes. The scale bar shows the branch length associated with 0.2 expected changes per site.

Several approaches were tried to improve resolution of the tree. Firstly, fast-evolving genes were excluded from the alignment. The resulting alignment, containing slowly-evolving genes, was analysed in three different ways: (**1**) excluding species with low numbers of characters; (**2**) including additional representatives for each class and (**3**) allowing independent branch lengths between genes.

To exclude fast-evolving genes from the analysis, evolutionary rates were calculated for each gene (the average logDet distance between pairs of taxa; Table 4.6). The 7 most slowly-evolving genes were in the new alignment (ACTIN, RNA_SSU, COX1, CYTB, COX2, ND1, EF1A).

Gene name	Average LogDet distance
ACTIN	0.3
RNA_SSU	0.41
COX1	0.64
CYTB	0.65
COX2	0.68
ND1	0.75
EF1A	0.79
COX3	0.8
H3	0.84
ND5	0.89
ND4	0.91
ND3	0.94
RNA_12S	0.94
RNA_16S	0.98
ATP6	1.01
RNA_LSU	1.06
ND2	1.11
ATP8	1.16
ND4L	1.17
ND6	1.38

Table 4.6: Average LogDet distances between taxa in the L1 dataset

The light grey box highlights those genes selected as slowly-evolving.

When the alignment of slowly-evolving genes was analysed without species with low numbers of characters present (*C. pamelae, H. depressa, Chaetoderma sp.* and *P. malayi*), the resulting tree (Figure 4.9) showed moderate support for a clade uniting molluscs and annelids (95% posterior probability). None of the three phyla were monophyletic and several unlikely relationships were found with high support. Polychaeta was sister taxon to Cephalopoda, making both molluscs and annelids paraphyletic. Clitellata was sister taxon to a clade containing Polyplacophora and Bivalvia. Turbellaria was sister taxon to all other ingroup taxa. Despite the selection of slowly-evolving genes used, some unbroken long branches were still present.

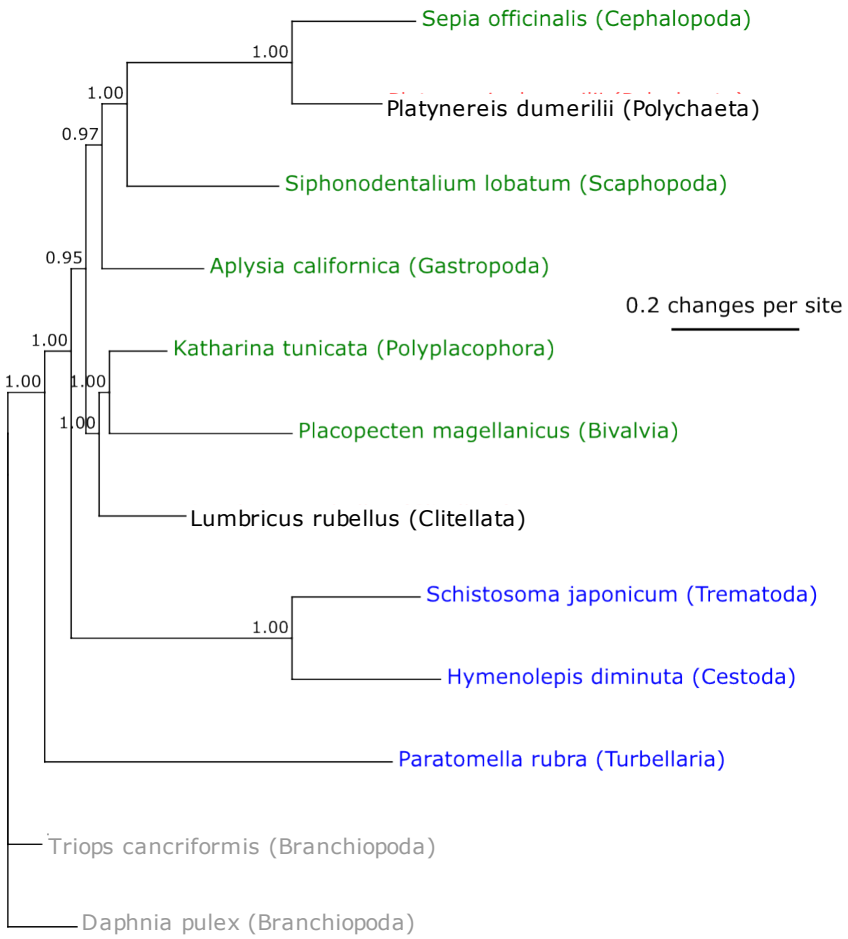


Figure 4.9: Tree derived from Bayesian analysis of slowly-evolving genes in well-represented taxa from the L1 dataset

Terminal nodes are labelled with the binomial name and the class in parentheses. Species are coloured according to phylum: black – Annelida, green – Mollusca, blue – Platyhelminthes. The scale bar shows the branch length associated with 0.2 expected changes per site.

In contrast, when the analysis was repeated with an additional representative of each class included, the tree gave a clearer picture of lophotrochozoan relationships. The

L1 dataset with additional taxa is shown in Table 4.7 and the tree in Figure 4.10).

NCBI taxid	Phylum	Class	Species	Characters
147105	Annelida	Branchiobdellae	*Cambrincola pamelae*	3051
55823	Annelida	Branchiobdellae	*Xironogiton victoriensis*	2522
35632	Annelida	clitellata	*Lumbricus rubellus*	18639
6398	Annelida	Clitellata	*Lumbricus terrestris*	15868
6421	Annelida	Hirudinida	*Hirudo medicinalis*	6753
279730	Annelida	Hirudinida	*Haementeria depressa*	6976
6359	Annelida	Polychaeta	*Platynereis dumerilii*	15005
195264	Annelida	Polychaeta	*Orbinia latreillii*	14102
211850	Mollusca	Aplacophora	*Helicoradomenia sp.*	4766
211838	Mollusca	Aplacophora	*Chaetoderma sp.*	4772
6565	Mollusca	Bivalvia	*Crassostrea virginica*	18684
6577	Mollusca	Bivalvia	*Plactopecten magellanicus*	20148
34570	Mollusca	Cephalopoda	*Sepioteuthis lessoniana*	16594
6610	Mollusca	Cephalopoda	*Sepia officinalis*	16868
6500	Mollusca	Gastropoda	*Aplysia californica*	19998
34582	Mollusca	Gastropoda	*Ilyanassa obsoleta*	18863
34587	Mollusca	Polyplacophora	*Katerina tunicata*	14607
58794	Mollusca	Polyplacophora	*Chaetopleura apiculata*	7761
55752	Mollusca	Scaphopoda	*Graptacme eborea*	12182
203167	Mollusca	Scaphopoda	*Siphonodentalium lobatum*	12251
6210	Platyhelminthes	Cestoda	*Echinococcus granulosus*	15129
6216	Platyhelminthes	Cestoda	*Hymenolepis diminuta*	18749
82849	Platyhelminthes	Monogenea	*Diclidophora denticulata*	5792
54594	Platyhelminthes	Monogenea	*Polystomoides malayi*	6510
6182	Platyhelminthes	Trematodda	*Schistosoma japonicum*	22517
6183	Platyhelminthes	Trematodda	*Schistosoma mansoni*	20091
90914	Platyhelminthes	Turbellaria	*Paratomella rubra*	12699
6161	Platyhelminthes	Turbellaria	*Dugesia japonica*	10088

Table 4.7: Details of species included in the L1 dataset with additional species

Columns give the NCBI taxonomic id, the phylum and class to which the species belongs, the binomial name and the total number of nucleotide characters in all genes. Species were automatically selected by TaxMan such that each class is represented by the species with the most complete set of sequences.

Molluscs and annelids were grouped to the exclusion of platyhelminths with very strong support (pp=0.98). Annelids were monophyletic with the exception of *P. dumerilii*, with Polychaeta sister taxon to the other classes. Classes within Annelida

were monophyletic with the exception of Polychaeta. Mollusca was paraphyletic, although three mollusc classes – Gastropoda, Cephalopoda and Polyplacophora – were monophyletic. Bivalvia was paraphyletic due to the inclusion of an aplacophoran. Platyhelminthes was monophyletic with the exception of the acoel *P. rubra*, whose position was poorly supported. Turbellaria was sister taxon to the parasitic classes. The three parasitic platyhelminth classes – Trematoda, Cestoda and Monogenea – were monophyletic.

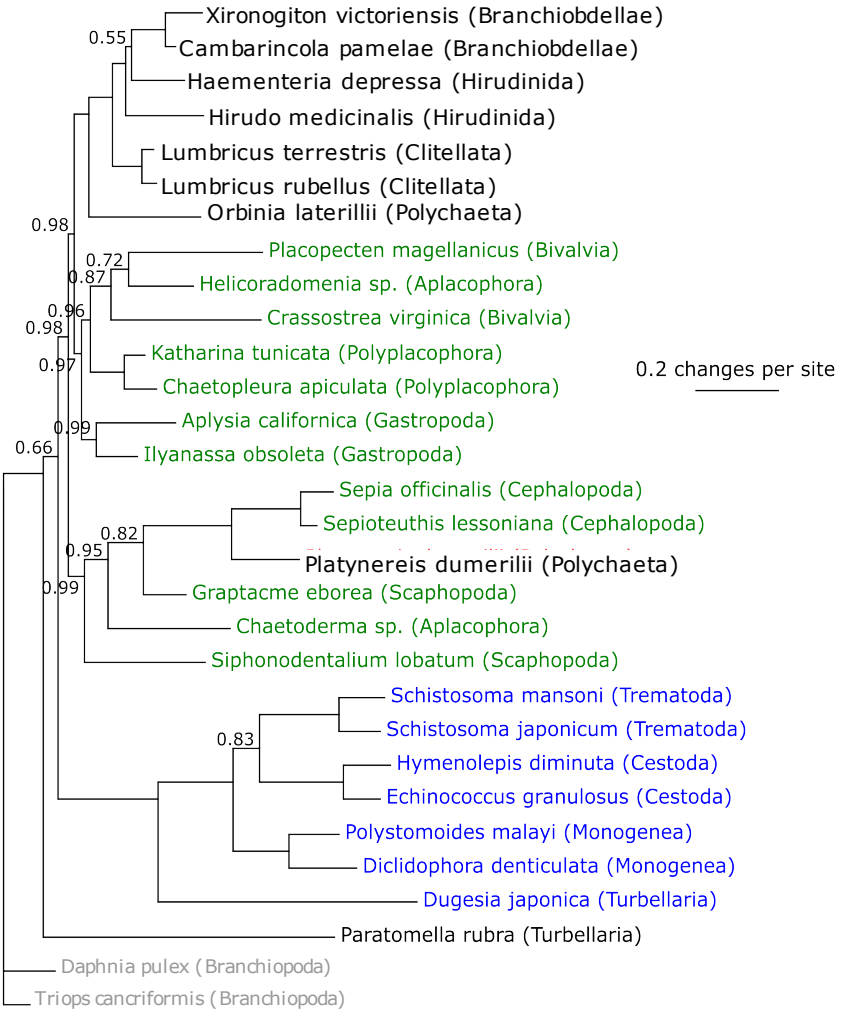


Figure 4.10: Tree derived from Bayesian analysis of combined genes from the L1 dataset with additional taxa

Terminal nodes are labelled with the binomial name and the class in parentheses. Species are coloured according to phylum: black – Annelida, green – Mollusca, blue – Platyhelminthes. The scale bar shows the branch length associated with 0.2 expected changes per site. Branches are labelled with posterior probability where <1.00

To investigate the presence of heterotachy among the slowly-evolving genes, the analysis of slowly-evolving genes for well-represented species **(1)** was repeated with branch lengths unlinked across genes. The resulting cladogram is shown in Figure 4.11 and the phylograms for individual genes are shown in Figure 4.12.

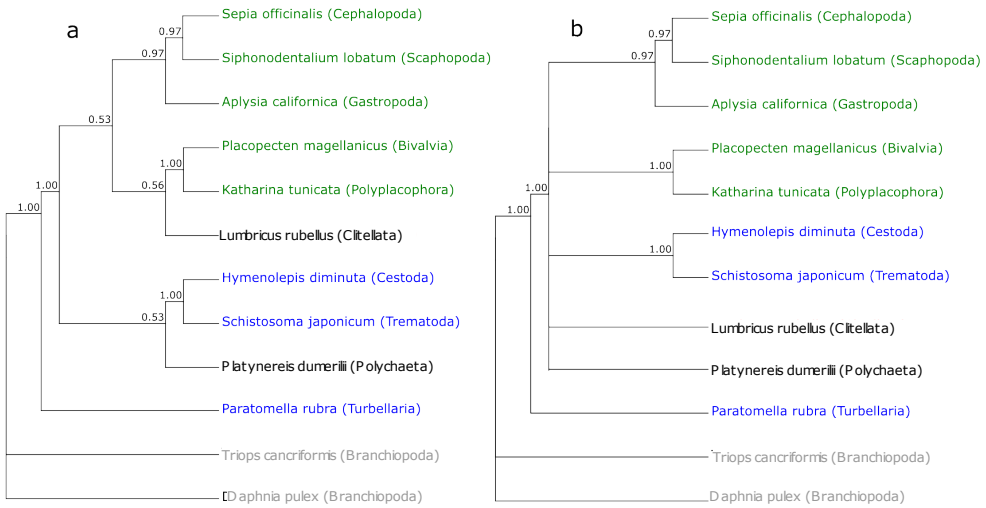


Figure 4.11: Cladograms derived from Bayesian analysis of slowly evolving genes for well-represented taxa from the L1 dataset. Branch lengths were unlinked across genes.

In (a), all branches are shown. In (b), only branches with >90% posterior probability are shown.

Terminal nodes are labelled with the binomial name and the class in parentheses. Species are coloured according to phylum: black – Annelida, green – Mollusca, blue – Platyhelminthes. The scale bar shows the branch length associated with 0.2 expected changes per site.

The tree resulting from this analysis shows only one robustly supported surprising placement, that of Turbellaria as a sister taxon to the other ingroup taxa. Other relationships are unresolved or poorly supported, with the exception of Cestoda + Trematoda, Bivalvia+Polyplacophora, and (Cephalopoda+Scaphopoda)+Gastropoda.

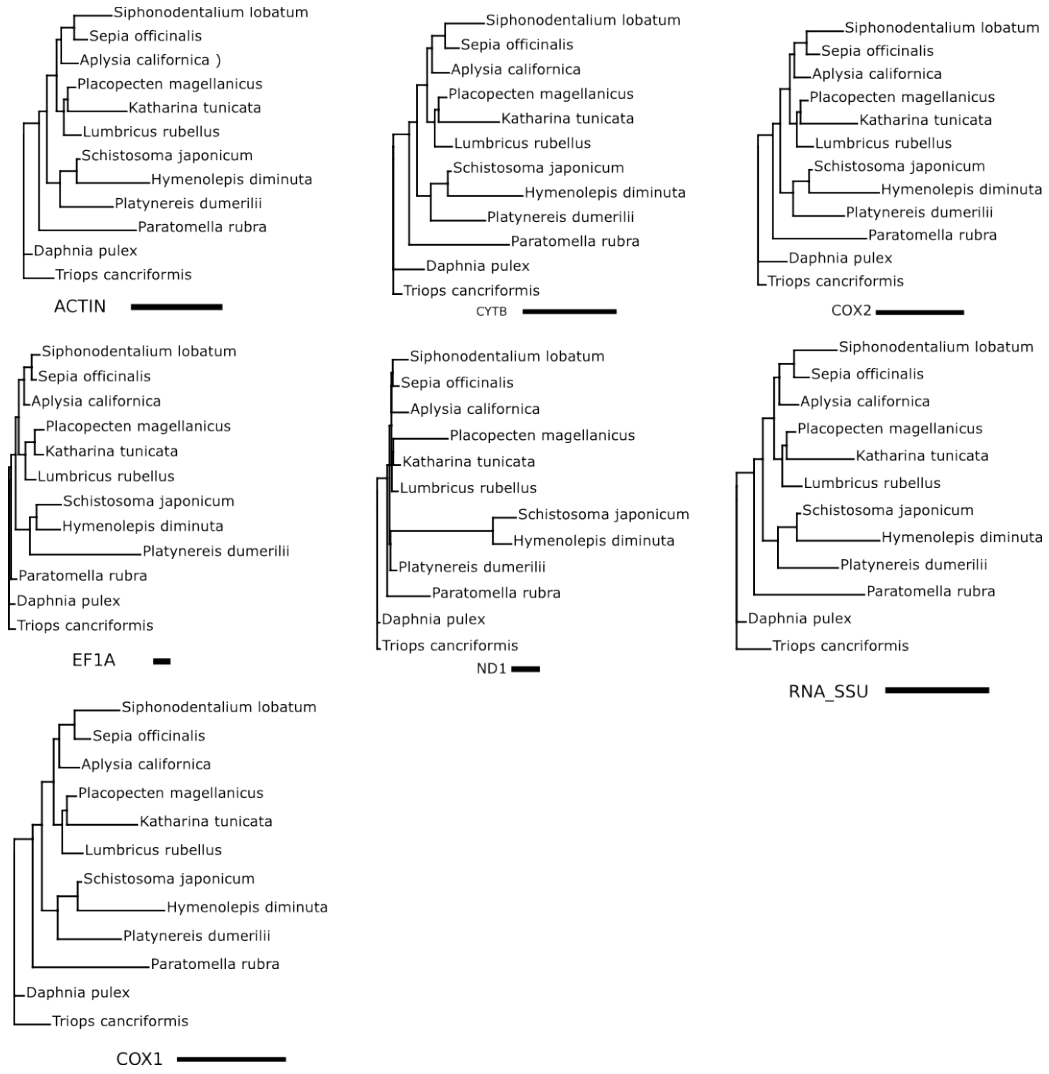


Figure 4.12: phylograms showing branch lengths for individual genes in Bayesian analysis of slow-evolving genes from the L1 dataset

Each tree has the same topology and support values as that in Figure 4.11. Bayesian analysis has been used to allow different genes to have independent branch lengths. Each tree is labelled with the gene it represents, and a scale bar showing the branch length associated with 0.2 expected changes per site.

Examining the branch lengths allocated to different genes, it is clear that while the genes have been selected for a slow rate of evolution, individual taxa have long branches leading to them. Additionally, a degree of heterotachy is present in the dataset: i.e. some taxa have long branches for subsets of genes. *P. dumerilli* has a much longer branch in EF1A than it does in other genes. In ND1, the two parasitic platyhelminth taxa, *S. japonicum* and *H. diminuta*, have long branches relative to other taxa, whereas in other genes only *H. diminuta* has a long branch. The better fit of the model with independent branch length is confirmed by Bayes Factor analysis. The Bayes Factor in favour of the more complex model, estimated as twice the difference in log-likelihood between the two runs, was 1296.

4.4.3 L2 dataset - neglected phyla

Table 4.8 gives details of the species included in the L2 dataset.

NCBI taxid	Phylum	Class	*Species*	Characters
60532	Acanthocephala	Palaeacanthocephala	*Leptorhynchoides thecatus*	16226
84287	Acanthocephala	Archiacanthocephala	*Oligacanthorhynchus tortuosa*	6232
178082	Acanthocephala	Polyacanthocephala	*Polyacanthorhynchus caballeroi*	6236
317552	Acanthocephala	Eoacanthocephala	*Neoechinorhynchus saginata*	5262
6359	Annelida	Polychaeta	*Platynereis dumerilii*	15001
35632	Annelida	clitellata	*Lumbricus rubellus*	18640
6216	Annelida	Cestoda	*Hymenolepis diminuta*	18747
279730	Annelida	Hirudinida	*Haementeria depressa*	6972
7574	Brachiopoda	Lingulata	*Lingula anatina*	14140
34513	Brachiopoda	Rhynchonellata	*Terebratalia transversa*	17721
67897	Brachiopoda	-	*Phoronis psammophila*	14329
33501	Brachiopoda	Craniata	*Neocrania anomala*	5106
231349	Bryozoa	Stenolaemata	*Crisia sp.*	4910
231348	Bryozoa	Gymnolaemata	*Bugula turrita*	6354
231373	Cycliophora	-	*Symbion sp.*	3455
6431	Echiura	-	*Urechis caupo*	18599
232741	Entoprocta	-	*Barentsia gracilis*	5202
6610	Mollusca	Cephalopoda	*Sepia officinalis*	16871
211838	Mollusca	Aplacophora	*Chaetoderma sp.*	4772
203167	Mollusca	Scaphopoda	*Siphonodentalium lobatum*	12250
6577	Mollusca	Bivalvia	*Plactopecten magellanicus*	20151
6500	Mollusca	Gastropoda	*Aplysia californica*	20003
34587	Mollusca	Polyplacophora	*Katerina tunicata*	14590
231370	Nemertea	Enopla	*Oerstedia dorsalis*	5774
6221	Nemertea	Anopla	*Cerebratulus lacteus*	6645
6182	Platyhelminthes	Trematoda	*Schistosoma japonicum*	22521
90914	Platyhelminthes	Turbellaria	*Paratomella rubra*	12694
54594	Platyhelminthes	Monogenea	*Polystomoides malayi*	6510
53701	Pogonophora	Perviata	*Galathealinum brachiosum*	8983
6426	Pogonophora	Vestimentifera	*Riftia pachyptila*	15300
104778	Rotifera	Seisonidea	*Seison nebaliae*	5307
96446	Rotifera	Monogononta	*Lecane bulla*	5772
96448	Rotifera	Bdelloidea	*Philodina roseola*	9709
6442	Sipuncula	Sipunculidea	*Phascolopsis gouldii*	12622
210783	Sipuncula	Phascolosomatidea	*Apionsoma misakianum*	5898
6669	Arthropoda	Branchiopoda	*Daphnia pulex*	19461
194544	Arthropoda	Branchiopoda	*Triops cancriformis*	17180

Table 4.8: Taxa included in the L2 dataset

Columns give the NCBI taxid, the phylum and class to which the species belongs, the species name and the total number of characters present for all genes

Becuase the L2 dataset included representatives of many more phyla than the L1 dataset, it included a much greater proportion of missing data. Taking into account the number of taxa with genes absent, and the reduction in the number of characters by selection of conserved positions using Gblocks, a combined analysis was necessary in order to test most of the phylogenetic hypotheses. For this combined analysis, the intention was to apply 'best practice' methods of phylogenetic reconstruction. To this end, the dataset was partitioned by gene, with independent model parameters for each gene. Gblocks was used to select well-conserved positions for analysis. 8,605 characters were analysed. Gblocks did not select any positions from the following genes: ATP8, ND2 ,ND3, ND4, ND4L, ND5 and RNA_16S. Mitochondrial protein-coding genes were recoded according to the NTE scheme while mitochondrial ribosomal RNA genes were recoded as purine/pyrimidine (R/Y). A GTR+G+I model was applied to all genes, with the exception of the R/Y coded mitochondrial RNA genes where a JC+G+I model was used. Third position bases were excluded from mitochondrial protein-coding genes. The resulting tree was well-resolved (Figures 4.13 and 4.14) and showed support for several of the relationships mentioned in the Introduction.

Annelida/Echiura/Pogonophora

As previously proposed (Giribet *et al.* 2000; Passamaneck and Halanych 2006), a clade was found uniting Annelida, Echiura and Pogonophora. Notably, this clade also included a monophyletic Sipuncula as a sister taxon to polychaete annelids, making Annelida paraphyletic. The mollusc class Aplacophora was sister taxon to a clade containing the two representatives of Pogonophora, though this taxon was represented

by very few characters.

Trochozoa/ Lophophorata

The Trochozoa hypothesis, uniting animals with a trochozoan larva, was not recovered. The clade uniting annelids, molluscs, sipunculans and nemerteans also contained brachiopods as a sister taxon to the Annelida/Echiura/Pogonophora + Sipuncula group. However, the tree did support the grouping of molluscs and annelids to the exclusion of platyhelminths. The Lophophorata hypothesis, uniting animals with a lophophorate feeding structure, was also not recovered. Bryozoans plus Turbellaria were found to be sister taxon to the remaining ingroup taxa.

Syndermata

A clade was found grouping rotifers and acanthocephalans with strong support. Within this clade, Acanthocephala was monophyletic and was sister taxon to the rotifer class Seisonidea (as found in Herlyn *et al.* [2003] with SSU data).

Cycliophora

Cycliophora was recovered as sister taxon to Entoprocta.

Platyzoa

Platyzoa, consisting in this taxon set of platyhelminths, rotifers and acanthocephalans, was not recovered, since platyhelminths were sister taxon to a clade containing (trochozoan taxa + Brachiopoda) + (Cycliophora + Entoprocta).

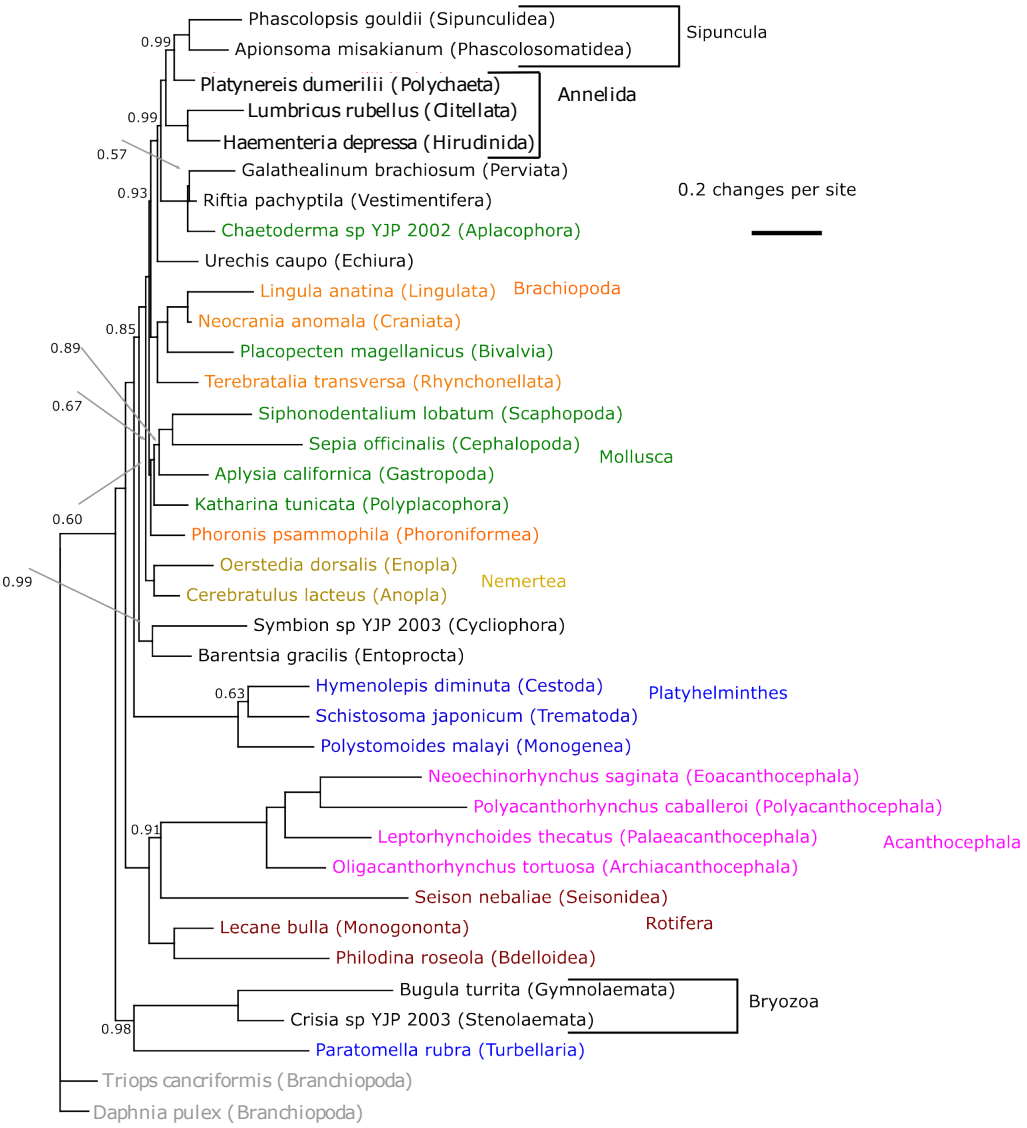


Figure 4.13: Tree derived from Bayesian analysis of combined genes in the L2 dataset

Colours indicate phylum membership. Branch posterior probabilities are shown when <1.00. Each species is labelled with the binomial name and the class in parentheses. The scale bar shown the branch length associated with 0.2 expected changes per site.

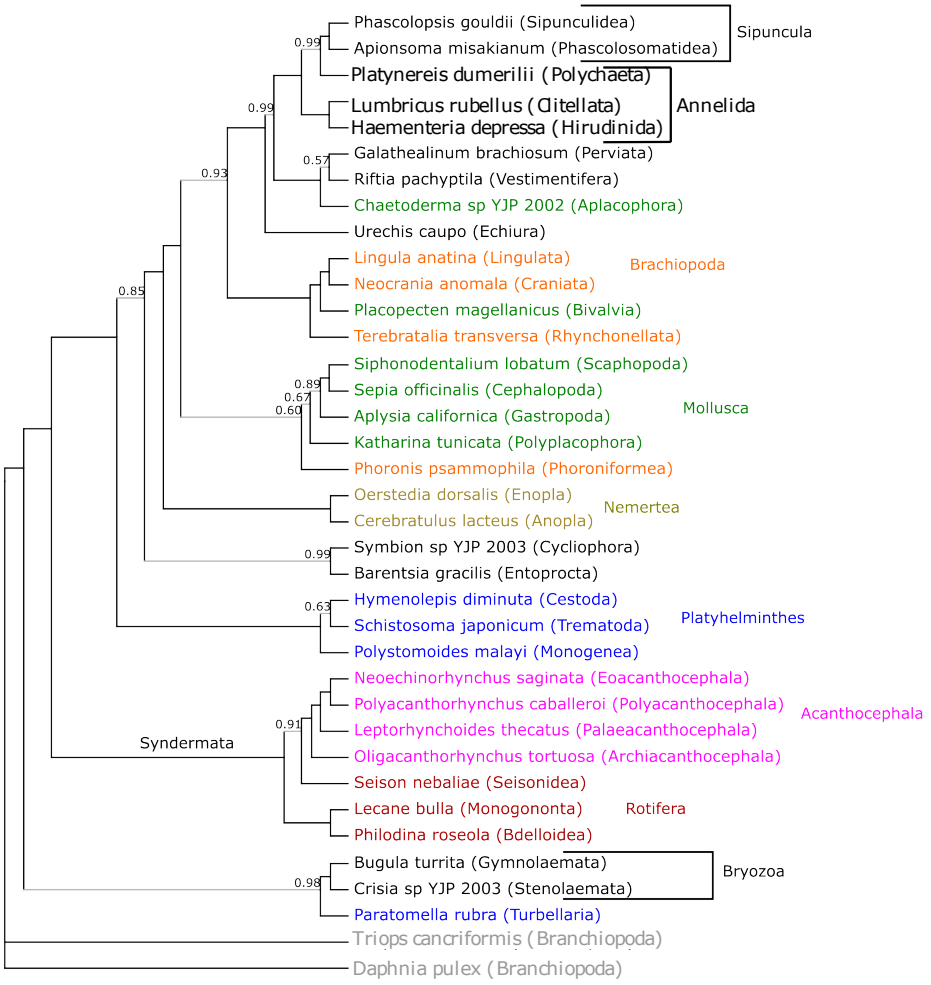


Figure 4.14: Cladogram derived from Bayesian analysis of combined genes from the L2 dataset

Colours indicate phylum membership. Bold branches have a posterior probability of 1.00; grey branches have a posterior probability <1.00 and are labelled. Each species is labelled with the binomial name and the class in parentheses.

The representative of Turbellaria chosen by TaxMan for the L2 dataset was *P. rubra*, an acoel flatworm. To test the effect of including a different turbellarian representative, the analysis was repeated with the inclusion of the triclad flatworm *Dugesia japonica*. The addition of this species to the alignment resulted in striking differences in the tree (4.15). While *P. rubra* remained near the base of the tree in a clade with the bryozoans, *D. japonica* was placed within Platyhelminthes as sister taxon to the parasitic classes. A clade containing Acanthocephala, Rotifera and Platyhelminthes (=**Platyzoa**) was now supported with a posterior probability of 0.96. Other relationships were largely unchanged, with the exception of the movement of two molluscs to near the base of the tree, and generally lower support values.

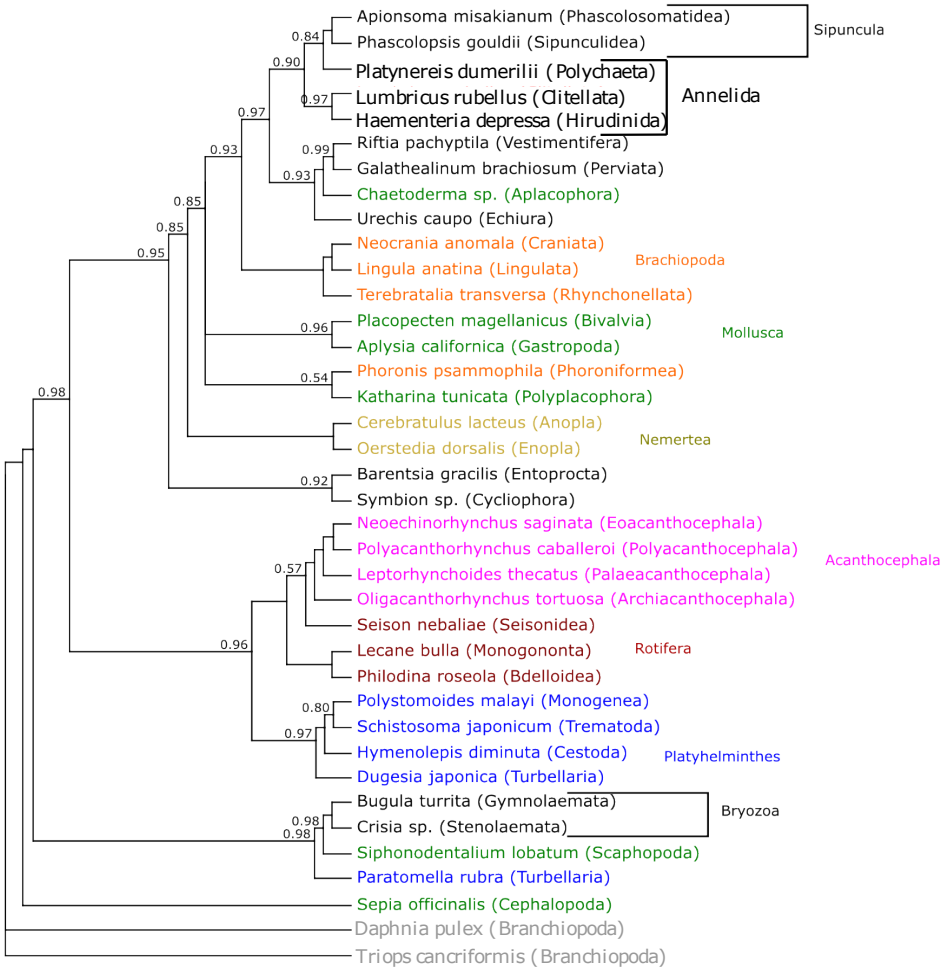


Figure 4.15: Tree derived from Bayesian analysis of combined genes from the L2 dataset with additional turbellarian

Branches are labelled with posterior probability if <1.00

4.5 Discussion

4.5.1 Phylogenetic methods and datasets

The two datasets examined here offer comment on a long-standing question in phylogenetics; is it better to include taxa with a significant proportion of missing data in order to improve the taxon sampling in an analysis (Kearney 2002; Rokas and Carroll 2005; Wiens 2003)? The L1 dataset contained mostly well-studied taxa while the L2 dataset included representatives from less well-studied phyla with less sequence data. The results of the L1 combined analysis including all genes underline the difficulties of resolving deep relationships between taxa. Despite the taxa included in the combined analyses (Figures 4.8 to 4.11) having 17,535 conserved aligned positions selected by Gblocks, only the analysis with additional taxa showed clear evidence in favour of one of the hypotheses of lophotrochozoan relationships. In the combined analysis of the slowly-evolving genes for the L1 dataset, different approaches were tried to improve the resolution of the tree: (**1**) excluding species with low numbers of characters; (**2**) including additional representatives for each class and (3) allowing independent branch lengths between genes. Excluding species with low numbers of characters (**1**, Figure 4.9) decreased the amount of missing data, but also decreased the taxon sampling, and lead to a tree that had high support values but proposed untenable relationships. In contrast, including an additional representative for each class (**2**, Figure 4.10) increased the amount of missing data, but lead to a tree with more plausible relationships which were very strongly supported. Finally, allowing independent branch lengths between genes (**3**, Figure 4.11) lead to a tree with low

support and very low resolution. The lack of resolution in the tree could be explained by two factors. The signal present in the original tree (Figure 4.8) could be due to model mis-specification, in which case the less resolved tree, found under the more accurate model, is better. Alternatively, the increase in the number of parameters (number of taxa = 12, number of partitions = 7, 126 additional branch length parameters) may have decreased the accuracy of parameter estimation, resulting in more variable trees and hence lower resolution. It would be informative to use this more complex model to analyse an alignment containing more genes (not just the slowly-evolving genes) and more taxa (e.g. with multiple representatives per class, see **2**) but such an analysis is not computationally feasible at this time (see Heterotachy, below). Overall, it seems that the most effective strategy to increase resolution was to add additional taxa to improve taxon sampling (**2**). This is in concordance with the results obtained in the L2 dataset, where inclusion of partially-represented species has given a tree with robust support, resolving relationships between both well-studied and neglected taxa. One might expect improved taxon sampling to be an important factor more often for very deep phylogeny, when the evolutionary distances between taxa of interest are very large and hence branches are long in absolute terms, than for phylogeny at lower levels. Adding taxa to 'break up' long branches is a common strategy in phylogenetics (Graybeal 1998; Slack *et al.* 2006; Wiens 2005) and in this case it appears to have been successful. While clearly useful, inclusion of additional taxa in phylogenetic analyses increases the computational complexity and is therefore a mixed blessing in the sense that a more complete dataset (in terms of taxonomic sampling) may not be suitable for some types of analysis. In particular, analysis of the

L1 combined dataset under an independent-branch-length model (Figure 4.11) did not recover many of the branches that were strongly supported under a simpler model. I cannot rule out the possibility that relationships supported by analysis of the combined L2 model (Figure 4.14) would similarly disappear under an independent-branch-lengths model. Such an analysis was not carried out due to computational constraints.

It seems likely that, at the phylum or class level, single representatives will not offer an accurate picture, and broader taxon sampling must be employed to ensure that the full diversity of target groups is included. The issue is complicated by the extremely large evolutionary distance between sister taxa. It seems likely that species with missing data may be most accurately placed when a closely related species will full sequence representation is also present in the tree. For diverse phyla, species from different classes may be separated by too great an evolutionary distance for this phenomenon to occur. A notable feature of all the analyses involving the L1 taxa is the sensitivity of the phylogenetic conclusions to taxon choice. For most trees, choosing any single mollusc (for example) as the exemplar of that phylum would give different conclusions depending on the species chosen. This underlines the importance of adequate taxon sampling in large diverse groups, and reminds us that any single species may not be a good representative of its taxon for phylogenetic purposes. The automated sequence-gathering and taxon-selection approach implemented in TaxMan allowed alternative taxon sets and extra species to be analysed very rapidly, making it easier to address these problems.

Heterotachy

A barrier to straightforward analysis of this question is the heterotachy seen between gene groups (and presumably between genes within a group). For instance, the acoel turbellarian *P. rubra* has a long branch leading to it for nuclear RNA genes and (to a lesser extent) mitochondrial protein-coding genes, but not for mitochondrial RNA genes or nuclear protein-coding genes. Similarly, the three parasitic platyhelminth classes have long branches for mitochondrial protein-coding genes but much less so for nuclear RNA genes. This type of variation in evolutionary rate across lineages for a single gene group or gene is not accommodated by evolutionary models which specify a single set of branch lengths. The use of a rate multiplier in MrBayes allows for different genes to be 'fast' or 'slow' and a non-clocklike tree allows different lineages to be 'fast' or 'slow'. However, such models do not allow a gene to be 'fast' in one lineage and 'slow' in another. Similarly, they do not allow a lineage (i.e. a branch) to be 'fast' for one gene and 'slow' for another. Use of a model including independent sets of branch lengths (but a common topology) across genes addresses this problem, but is extremely computationally intensive and is unlikely to be feasible for large numbers of taxa. For n taxa, a fully resolved unrooted tree has $2n$-3 branches; therefore each partition adds an extra $2n$-3 parameters to the analysis. However, it is likely that current Bayesian inference software is not optimised for this type of analysis, and that future development might yield performance gains. The results of the Bayes Factor analysis of models with and without independent branch lengths between genes shows the extent of the problem. According to the interpretation of Kass and Raftery (1995), a BF of >20 is considered strong evidence for the favoured model;

therefore the value of >1200 found in favour of the model with independent branch lengths indicates that it explains the data much better than the alternative model with a single set of branch lengths. The fact that support for the majority of relationships disappears under this model suggests that they may have been an artefact of model mis-specification.

4.5.2 Molluscs, annelids and platyhelminths

As well as supporting the grouping of molluscs and annelids to the exclusion of platyhelminths, the results reported here suggest some interesting relationships within phyla. Evidence for grouping of the parasitic platyhelminth classes to the exclusion of the free-living class was found in trees derived from nuclear RNA genes, mitochondrial RNA genes and mitochondrial protein-coding genes in the L1 dataset, and in combined analysis of the L2 dataset. These analyses were carried out under the assumption that Platyhelminthes (and, within it, Turbellaria) was monophyletic, as specified by the NCBI taxonomy. The results of the L2 analysis with an additional turbellarian flatworm suggests that the acoel flatworms, represented by *P. rubra*, should not be placed within the Platyhelminthes. This result leaves two hypotheses open. If acoels are lophotrochozoans, then they are placed near the base of that group as shown in Figure 4.14, possibly allied with bryozoans. Alternatively, if acoels are not lophotrochozoans but basal triplobasts, the analyses with an ecdysozoan outgroup described in this chapter are insufficient to place them. In this second scenario, acoels should form the outgroup in Figure 4.14 and in the summary cladogram below. This result was found in an analysis of 18S ribosomal RNA and morphology for

invertebrate taxa, in which acoel flatworms were found to be sister taxon to all other triploblastic animals (Littlewood, Rohde and Clough 1999). These two hypotheses could be tested in an analysis including lophotrochozoan, ecdysozoan and deuterostome phyla and a non-triplobast outgroup.

The results from the L2 dataset underline the importance of comprehensive taxon sampling in deep phylogeny, and strongly suggest that the inclusion of the acoel flatworms in Platyhelminthes is incorrect. The placement of the non-acoel turbellarian *D. japonica* is congruent with the idea of an ancestral free-living platyhelminth and a single acquisition of the parasitic lifestyle in flatworm evolution, as suggested by Littlewood, Rohde and Clough (1999). In that work, the authors find the turbellarians to be paraphyletic, with the parasitic flatworms (Neodermata) arising from within the Turbellaria. This hypothesis was not tested in the current work; however, ample sequence data exist to address questions of platyhelminth evolution with a multigene approach.

Within the annelids, all four gene groups of the L1 dataset support a division between polychaetes and the remaining three classes. Combined analysis of the L2 dataset also supports this, although Branchiobdellae is not represented and Annelida is paraphyletic. This is in agreement with non-molecular synapomorphies for polychaetes (parapodia, elaboration of the head) and for the other classes (hermaphroditism).

Notably, in the single-group L1 analysis and in the combined L2 analysis molluscs, annelids and platyhelminths were consistently paraphyletic. To investigate this result would require increased taxonomic sampling from within the various classes to identify sources of bias.

4.5.3 Neglected lophotrochozoan phyla

The high degree of resolution seen in the combined analyses of the L2 dataset allows us to infer support, or lack of support, for a number of phylogenetic hypotheses. Table 4.9 reproduces the information shown in Table 4.2, with this work included.

		Conclusions						
Citation	**Characters**	**Lophophorata**	**Platyzoa**	**Cycliophora + Entoprocta**	**Syndermata**	**Trochozoa**	**Annelida + Pogonophora + Echiura**	**Platyhelminthes**
Garey et al. 1996	18S ribosomal RNA	-	-	-	monophyletic[1]	-	-	-
Giribet et al. 2000	18S ribosomal RNA, 276 morphological characters	paraphyletic	monophyletic[2]	paraphyletic[3]	monophyletic	monophyletic	monophyletic	monophletic
Peterson and Eernisse 2001	18S ribosomal RNA, 138 morphological characters	monophyletic	-	-	-	monophyletic	monophyletic	not monophyletic[4]
Herlyn et al. 2003	18S ribosomal RNA	-	-	-	monophyletic[5]	-	-	-
Anderson, Cordoba and Thollesson 2004	Na-K ATPase a	paraphyletic	-	-	-	-	-	-
Passamaneck and Halanych 2006	18S ribosomal RNA, 28S ribosomal RNA	paraphyletic	monophyletic	monophyletic	monophyletic	paraphyletic	monophyletic	-
This work	12S, 16S, 18S, 28S ribosomal RNA, ACTIN, H3, EF1A, 12 mitochondrial protein coding genes	paraphyletic	monophyletic	monophyletic	monophyletic[5]	paraphyletic[6]	monophyletic[7]	not monophyletic[4]

1 – Acanthocephala sister taxon to Bdelliodea
2 – With Gastrotricha and Gnathostomulida and Cycliophora
3 – Cycliophora sister taxon to Syndermata
4 – Acoel flatworms not included in Platyhelminthes
5 – Acanthocephala sister taxon to Seisonidea
6 – Brachiopods group within Trochozoa
7 – Sipuncula included in this group

Table 4.9: Summary of previous lophotrochozoan analyses with this work included

Columns give the citation, characters used in the analyses, and conclusions regarding the status of each hypothesis. A hyphen (–) indicates that the study did not address this question, or that there was no support to evaluate it.

The **Trochozoa** hypothesis unites animals with a trochozoan larva – molluscs, annelids, nemerteans and sipunculids. The L2 tree groups these phyla along with brachiopods, suggesting either that the trochozoan larval condition has evolved multiple times, or that it has been lost in brachiopods. This branch also falsifies the **Lophophorata** hypothesis which unites brachiopods and bryozoans on the basis of the

lophophore. Previous studies have also found this result (Passamaneck and Halanych 2006), suggesting convergent evolution of lophophore-like organs. In the L2 analysis with an additional turbellarian species, **Platyzoa** was recovered as a sister taxon to the Trochozoa+Brachiopoda clade mentioned above. Gastrotricha was not included in the analysis, so its inclusion in Platyzoa was not tested.

As suggested by several workers (reviewed in Halanych [2004]) **annelids**, **pogonophorans** and **echiurans** were found to be closely related. This relationship, while supported in the L2 combined tree, was complicated by (1) the inclusion of Aplacophora as a sister taxon to Pogonophora and (2) placement of sipunculans as sister taxon to polychaete annelids. Relationships within this clade are in a state of flux (Halanych [2004]) and it seems prudent to wait for more directed studies before reaching conclusions. The **Syndermata** hypothesis, uniting rotifers and acanthocephalans, was strongly supported in analysis of the L2 dataset. Furthermore, Acanthocephala was placed within Rotifera rather than as a sister taxon, as previously found with SSU data (Garey *et al.* 1996), although this study suggested a sister-taxon relationship between Acanthocephala and Bdelloidea, rather than Seisonidea as was found here). A sister taxon relationship between **Cycliophora** and Entoprocta has been proposed on the basis of morphology (Funch and Kristensen 1995) and combined SSU and LSU data (Passamaneck and Halanych 2006) and was strongly supported in the combined L2 tree. The placement of Bryozoa as sister taxon to other ingroup taxa was found, with much lower support, in a previous analysis (Passamaneck and Halanych 2006) and was found in the L2 tree with strong support. However, both bryozoan species had a relatively small number of characters, and this result needs to

be confirmed with more data. **Acoelomorpha** did not group within the Platyhelminthes but was placed near the base of the tree. This is in contrast to its placement in the NCBI taxonomy and is evidence that acoel flatworms should be placed (1) at the base of the Lophotrochozoa or (2) outside Lophotrochozoa.

Taking into account the uncertainties in the trees, I advance the hypotheses of lophotrochozoan relationships shown in a summary cladogram (Figure 4.16). Where phyla were paraphyletic in the various analyses, I have placed them in the position suggested by the majority of representatives.

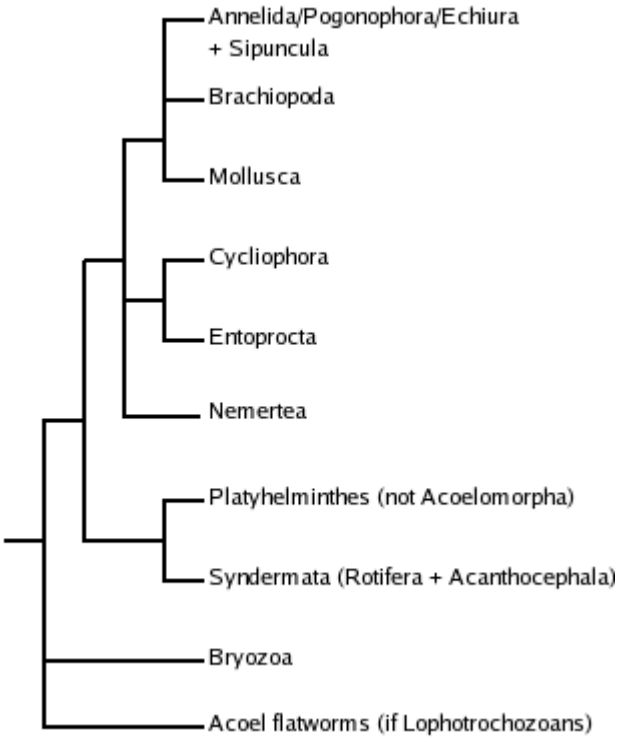


Figure 4.16: Summary cladogram of lophotrochozoan relationships

5 Summary discussion

5.1 Problems and solutions in multigene phylogenetics

In this thesis I have described the phylogenetic analysis of two large multigene datasets and discussed some of the problems that have arisen. Several of these issues have been encountered in previously published multigene studies and are likely to play a more and more prominent role as these types of analyses become more common. I have also presented a software tool, TaxMan, that assists in assembling datasets for phylogenetic analysis.

Accurate models for multiple genes

A common theme in both sets of analyses described in this thesis is the importance of using evolutionary models that take into account differences between genes. The importance of model choice in likelihood-based phylogenetic inference has been highlighted in many empirical and simulation studies (Lemmon and Moriarty 2004; Nylander *et al.* 2004; Brandley, Schmitz and Reeder 2005; Pupko *et al.* 2002). Models that assume a single model of evolution across all aligned sites will be inadequate if different genes evolve according to different patterns. The solution is to use a partitioned models, in which characters are allocated into sets, each of which can have independent model parameters. In the analyses cited above, partitioned models were found to outperform unpartitioned models when analysing heterogeneous data. Characters can be partitioned in many ways. Partitions can represent different genes or

groups of genes, or different codon positions within genes. For RNA data, partitions can represent stem and loop regions. Recently, methods have been developed to use Monte Carlo sampling to explore the range of possible partitions (Pagel and Meade 2004). Partitioned models are implemented in MrBayes 3 (Ronquist and Huelsenbeck 2003) using a scheme that allows each parameter to be linked or independent across any combination of partition. Partitions can have the same model (e.g. GTR+G+I) but independent model parameters. Typically, state frequencies, transition rates, gamma rate variation parameters and proportions of invariant sites are unlinked across partitions. However, MrBayes also permits partitions to have different overall rates (using a rate multiplier parameter), independent branch lengths (allowing for heterotachy) and even different models between partitions (e.g. GTR versus JC).

In the analysis of chelicerate orders (Chapter 3), partitioning the dataset to allow each gene to have different model parameters radically changed the conclusions. In the analysis of lophotrochozoan phyla (Chapter 4), allowing independent sets of branch lengths on a common topology between genes revealed that some well-supported relationships were questionable. In both these cases, Bayes Factor support for the more complex, partitioned model was overwhelming. While unlinking of model parameters is becoming more common when working with multiple genes (Brandley, Schmitz and Reeder 2005; Nylander *et al.* 2004), unlinking of branch lengths is not. Discussions of heterotachy in the literature have been confined to evaluating the performance of different tree reconstruction methods (Kolaczkowski and Thornton 2004; Philippe *et al.* 2005) rather than comparison of different models. Use of realistic models is crucial for robust phylogenetic analysis: when an alignment contains

multiple genes, realistic models are very likely to involve partitions. Further simulation studies, explicitly addressing the effects of (1) unlinking different combinations of parameters and (2) unlinking branch lengths would be a useful contribution to multigene phylogenetics.

Systematic bias in large datasets

I define bias, in the phylogenetic context, as any sequence characteristic that leads to incorrect phylogenetic conclusions. Rapid rates of evolution (Brinkmann *et al*. 2005), base composition (Foster and Hickey 1999) and mitochondrial strand bias (Hassanin 2006) all fall under this heading. In multigene phylogenetics, we are particularly interested in cases of bias that affects all characters in a genome equally, known as systematic bias. Increasing the number of characters available for analysis will not overcome systematic bias, since all characters are equally affected. In fact, confidence in the incorrect phylogenetic conclusion will increase with the number of characters under systematic bias making multigene studies particularly vulnerable.

Taking the example of evolutionary rate, examples of both systematic and non-systematic bias are known. It is well-known that inclusion of multiple sequences with rapid evolutionary rates (and hence long branches) in a phylogenetic analysis can cause long branch attraction (LBA; Felsenstein 1978, Brinkmann *et al*. 2005). This phenomenon takes place when independent but convergent changes in sequences with rapid evolution are misconstrued as shared derived changes (synapomorphies), leading tree reconstruction methods to group such taxa together.

Consider an alignment that includes two sequences from distantly-related taxa with long branches which are found to group together in phylogenetic analysis due to LBA.

If additional characters are added to the alignment, any change in the tree will depend on whether the rapid evolutionary rate of the long-branch sequences is characteristic of the species' genomes as a whole (i.e. systematic). In some lineages, single genes have undergone accelerated evolution relative to the rest of the genome (Kouprina *et al.* 2004). If this is the case for the gene represented in our conceptual alignment, then any additional characters (from other genes) added to the alignment will not be likewise affected and will therefore contribute genuine phylogenetic signal, which will eventually drown out the false signal contributed by the rapidly-evolving gene. The bias in this case is not systematic. However, some lineages are known to have an overall rapid rate of evolution which affects all genes (Brinkmann *et al*. 2005). If this is the case, then the bias is systematic, and any additional characters added to the alignment will be affected and will contribute the same erroneous phylogenetic signal. We might expect this scenario to be more common in the case of deep phylogenetic analyses, in which the terminal taxa are separated by large evolutionary distances, than in analyses involving closely-related species, simple because the former is more likely to contain lineages with very different overall rates of evolution.

Similar argument cans be made for other types of bias. In the case of mitochondrial strand bias (discussed in Section 3.2.3), taxa with similar patterns of bias will tend to be grouped together under phylogenetic analysis. This was described in detail for arthropods in Hassanin (2006). Mitochondrial strand bias is an interesting case because it affects only mitochondrial genes: it could therefore be said to be systematic, but only for a particular set of genes. From a practical point of view, the degree to which mitochondrial strand-bias is systematic depends on the dataset; in an analysis of

nuclear genes its effects will not be relevant. In the case of AT content, taxa with high (or low) AT content will tend to be grouped together. High or low AT content has generally been found to characterise entire genomes (Knight, Freeland and Landweber 2001), so it is likely that AT content will be a systematic bias in most cases. The presence of isochores in mammal and bird genomes (Bernardi et al. 1985) may complicate the picture.

In cases where large amounts of sequence data are available, known systematic bias can be overcome by careful character selection. In a recent example (Philippe, Lartillot and Brinkmann 2005), an LBA artefact was eliminated by excluding rapidly-evolving genes from a large dataset. This approach was only possible because of the very large amount of sequence data available – the artefact was not eliminated until 75 out of 146 genes had been discarded. A similar approach was used by Dopazo and Dopazo (2005). While this approach is currently only possible for the groups with the most sequence data, it illustrates how systematic bias, if recognised, can be tested for and managed. Similar approaches could be used to mitigate the effects of other types of bias by removing the characters most affected. For example, the neutral transitions excluded (NTE) recoding scheme described by Hassanin, Leger and Deutsch (2005), and used to analyse chelicerate relationships in Chapter 3, specifically recodes characters most likely to be affected by strand-bias. Although this approach acts on individual characters rather than entire genes, the principle is similar. Careful selection of taxa to represent higher groups can also be used to ameliorate the effects of systematic bias, provided (1) the systematic bias does not affect all species in the group and (2) there is an unaffected representative of the group with sufficient

sequence data. These circumstances are unlikely to hold for the majority of analyses.

Taxon selection

The automated approach to sequence acquisition results in assembly of a dataset containing many more species than can be included in a phylogenetic analysis (e.g. ~8,000 species for the Lophotrochozoa analysis described in Chapter 4). When seeking to resolve relationships between large taxonomic groups, individual species must be selected to serve as exemplars of that group. For any large group, the researcher must decide on the number of representative species, and on the criteria with which they are to be picked. The number of representative species that can be included in an analysis is limited by computational considerations and sequence availability. Increasing the number of representative species increases the accuracy and robustness of phylogenetic conclusions (Chapter 4; Graybeal 1998; Slack *et al.* 2006; Wiens 2005; Wiens 2006). However, it also makes the analysis more challenging, and may force the use of less sophisticated techniques. Additionally, sequence distribution within any given group is likely to be uneven over species, so increasing the number of representative species will often increase the proportion of missing data in an alignment (though simulation studies indicate that adding even highly incomplete taxa can be helpful [Wiens 2005]). The use of an explicit criterion for species selection allows objective, automated selection of taxa for analysis. Of particular concern is the fact that 'model organisms', chosen for a set of traits that make them amenable to experimentation (short generation time, easy to look after, etc.) may not be the most suitable examples of their group for phylogenetic analysis. Nevertheless, if sequence completeness is used as a selection criterion, such organisms will commonly be

chosen. In current phylogenetic studies, a form of taxon selection is used where species are excluded on the criteria of long branches, aberrant base composition, etc., but this has not been put in an automated framework.

The analysis of lophotrochozoan phyla described in Chapter 4 furnishes us with several examples of the effects of taxon selection. When analysing groups of genes, important phyla (e.g. Mollusca) were not monophyletic. In these cases, it seems likely that choosing a single representative species for that phylum would yield a phylogenetic conclusion dependent on the placement of that species, which may not be indicative of the placement of the phylum as a whole. Choosing multiple representatives from each phylum revealed the conflicting relationships suggested by each individual species and prevented incorrect conclusions being drawn.

When analysing a number of combined genes, the analysis that included only the best-represented phyla was inconclusive and proposed several relationships which did not fit any prior hypothesis, including paraphly of all three represented phyla. In dramatic contrast, the analysis that also included representatives of neglected phyla proposed a clear, believable set of relationships that confirmed several prior hypotheses. Clearly, in this case sequence completeness was not a good criteria for taxon selection.

The final example concerns the selection of an acoel flatworm as a representative of the free-living flatworms. Under phylogenetic analysis, this species did not group with the other flatworms. Subsequent analysis that used a different free-living representative revealed that the originally-chosen species was not a good representative as it was not truly closely related to other non-acoel flatworms.

There are parallels to be drawn between taxon selection and gene selection. In

phylogenetic analysis, we assume that a given gene sequence is representative of the genome of a species as a whole, at least as far as evolutionary history is concerned. By using multiple genes in a phylogenetic analysis we seek to minimise the effects of oddly-behaving genes. In the same way, when we select a species to represent a higher taxon we assume that it is representative of the taxon as a whole. By using multiple species to represent higher taxa we can recognise oddly-behaving species and avoid being mislead by their behaviour under phylogenetic analysis.

A common theme in the solution to these problems is the importance of exploring the different ways in which the dataset can be analysed. The effects of model choice on an analysis can be investigated by analysing the same alignment under different models and parameters. The alignment on which tree reconstruction is carried out can be changed between analyses, rather than being seen as a static entity. Systematic bias can be identified and corrected by changing the genes and characters used in an analysis. Issues of taxon sampling can be revealed by adding and removing taxa and re-analysing the alignment. The process of exploring a dataset in this way is iterative, with each analysis informing the next.

5.2 Tools for dataset exploration

Current phylogenetic tools are not well-suited to carrying out the type of analysis described above. In general, the way that sequence data are stored in public databases makes it difficult to explore the use of different genes and taxa in an alignment. Below

is a discussion of the various components that are required for a phylogenetic workbench and of the shortcomings of current offerings.

Phylogenetic databases

The current main repository for biological sequence data is the NCBI's GenBank and its various mirrors (Benson *et al.* 2006). While GenBank does incorporate a taxonomy, it is far from ideal for examining sequences in a phylogenetic context. The structure is inconsistent, with many species not assigned to intermediate taxonomic groups (order, family, etc.). Nodes representing species (taxids under the NCBI scheme) are occasionally renamed, leading to a mismatch between datasets downloaded on different dates. The taxonomy is inflexible, representing only a single view of evolution, and no straightforward mechanism exists for expressing alternative possibilities. A taxonomically aware database on the scale of GenBank that addressed the above issues would be a powerful tool for multigene phylogenetics and the need for such is widely accepted (Page 2005; Page and Valiente 2005). The Arthropod Mitochondrial Genome Accessible database (Feijao *et al.* 2006; discussed in Chapter 2) represents a step towards such a project. It allows taxonomic selection of orthologous genes but is restricted to a single reference taxonomy. By contrast, TreeBase (Sanderson *et al.* 1994) places more emphasis on storing trees and linking them through common taxa. Another shortcoming of the GenBank database for phylogenetics purposes is the lack of standards in gene annotation. Gene synonyms and different annotation standards between sequencing projects and communities mean that inferring orthologous relationships based on annotation is hard. A recent effort to quantify the degree of overlap of gene names confirmed these problems (Fundel and

Zimmer 2006). A database designed with phylogenetics in mind would implement standardised gene names and incorporate orthology groups into its design. It would also allow selection of taxa for analysis based on flexible criteria.

Alignment software

A crucial step in phylogenetic analysis is the alignment of multiple sequences. This is challenging in deep phylogenetics for two reasons: a large evolutionary distance between sequences, and the presence of missing data. Large evolutionary distance between sequences leads to lower sequence similarity and a more challenging alignment problem. Missing data, a characteristic of the types of analyses described in this thesis, can arise in one of two ways. Firstly, when sequences are obtained from public databases, some genes will not have been sequenced for a given species, in which case the entire gene will have to be coded as missing data. Secondly, a given gene may be represented by an EST sequence, in which case the sequence is likely to cover only a portion of the total gene length. The second scenario poses a particular problem for most multiple sequence alignment programs, which are not capable of carrying out the local alignment that is required for partial sequences. Other programs that deal with multiple sequence alignments can fail to take missing data into account. For example, Gblocks (Castresana 2000) is a program for automatic selection of conserved blocks from a multiple sequence alignment. Such a tool is very useful in deep phylogenetics, since it identifies characters for which character homology is more certain, and which are likely to be slowly-evolving. However, the software is currently of limited use, as it makes no distinction between gaps and missing data. A version of the software that recognised missing data would be able to identify more conserved

blocks and facilitate better phylogenetic analysis. Automatic evaluation of alignment quality would be a very useful step towards automating multigene analyses and some methods are already available (Ahola *et al.* 2006; Thompson *et al.* 2001) and have been used as part of some multigene phylogenetics workflows (Philippe et al. 2004) However, for such an approach to be more widely applicable it must take account of missing data and partial sequences.

Complex models with multiple partitions

When including multiple genes in a phylogenetic analysis, an evolutionary model must be used that takes into account the differences in patterns of evolution between genes. In a likelihood framework, this is normally described as a partitioned model, in which different genes are allocated to different partitions which are given independent model parameters. For multigene phylogenetics, tree reconstruction methods must support such parameter-rich models. Current software that does support complex models may not be optimised for them. For example, MrBayes (Ronquist and Huelsenbeck 2003) support models with many partitions, but the default rate multiplier proposal mechanism is inefficient at exploring parameter space when the number of partitions is large, causing analyses to fail to converge. To facilitate multigene analysis, programs that carry out tree reconstruction must be written with these types of analysis in mind, or promote best practice in parameter choices. Statistical testing of evolutionary models should also be a target application.

Orthology clustering

An area of key importance for multigene phylogenetics is *a priori* orthology assignment, in which collections of unidentified sequences from various species of

interest are clustered into orthology groups. Robust methods for doing this would open up vast amounts of sequence data to phylogenetic analysis, particularly from EST sequencing projects, where transcribed regions are cloned and sequenced randomly (leading to unidentified sequences). Early work on a number of methods seems promising (Alexeyenko *et al.* 2006; Chiu *et al.* 2006; Dufayard *et al.* 2005; Koonin 2005; Li, Stoeckert and Roos 2003; Tatusov *et al.* 2003), but confidence in their results remains low. Current solutions for orthology assignment implemented in TaxMan rely on a combination of annotation, similarity searching and *a priori* biological knowledge about suitable genes.

Testing for systematic bias

One recurring theme in this thesis has been the importance of systematic bias in multigene phylogenetics. Some workers have taken steps to develop methods for detecting and eliminating systematic bias – for example, Philippe, Lartillot and Brinkmann (2005) removed fast-evolving genes to reveal the effects of long branch attraction, and Hassanin (2006) implemented a recoding scheme to reveal the effects of mitochondrial strand-bias. These methods allow one to identify relationships that are caused by systematic bias, and to eliminate bias. The development of standard methods of testing for systematic bias would be a useful field of study.

5.3 A phylogenetic workbench

TaxMan (Chapter 2) was designed to address some of the issues described above and to facilitate phylogenetic analysis of large datasets. By (1) mining public sequence data to assemble a dataset of aligned orthologous genes and (2) making it easy to select

subsets of genes and taxa from that dataset for analysis, TaxMan encourages the type of exploration described earlier. TaxMan creates a database that fulfils some of the requirements outlined in the "Phylogenetic databases" section above. It standardises gene names to generate orthlogous sets of genes for analysis. It also carries out automatic selection of taxa for analysis. It attempts to overcome the problems outlined in the "Alignment" section by carrying out multiple sequence alignment at the protein level using a program that is capable of local alignment. It encourages the use of partitioned models by automatically dividing the alignment up into character sets for individual genes and codon positions.

Though TaxMan represents a useful step towards a comprehensive phylogenetic workbench, many improvements could be made, some of which have been discussed in Chapter 2. In the sequence gathering stage, future software should add support for a wider range of input formats, as well as allowing easy integration with local datasets. It should also use standalone orthology inference software to expand the amount of sequence that can be analysed. The user should be able to store multiple reference taxonomies, representing different phylogenetic hypotheses, and compare then to trees resulting from phylogenetic analysis in an automated fashion. Assessment of the sensitivity of phylogenetic conclusions to gene and taxon selection could also be automated, allowing sophisticated analyses to be scripted. In particular, users should be able to easily explore the effect of different taxon selection criteria. Future software should also place emphasis on statistical model testing, as work presented here and elsewhere has shown the importance of model choice in phylogenetic inference. With multigene phylogenetics in mind, testing for systematic bias will be an important task

for phylogenetic tools, and this should be reflected in the design. In order to test for systematic bias by systematically excluding characters, a high degree of automation will be necessary.

Because the computationally demanding parts of a phylogenetic study (alignment, tree-reconstruction, consensus building) are carried out by external programs, any major speed increases must come from those programs rather than from the workbench itself. However, future tools should make full use of well-developed technology, such as relational databases and parallelisation, to remain responsive when dealing with large datasets.

Widespread adoption of online databases has demonstrated the enormous potential for web-based tools to facilitate bioinformatic analyses. Future software should include components for generating web-accessible databases to allow public access to datasets of aligned sequences.

5.4 Best Practice

Working within TaxMan to analyse the two datasets presented in this thesis has lead to a number of considerations that constitute best practice in multigene phylogenetics that are backed up by the findings of previous multigene studies.

Use as much of the available data as possible

For most taxonomic groups, the amount of sequence data that has been used for phylogenetic analysis represents only a fraction of that which is available. Use of TaxMan for data gathering in Chapters 3 and 4 has demonstrated that previously

unused sequence data can be assembled and analysed using an automated approach. Recent studies have shown that using large numbers of genes benefits phylogenetic reconstruction (Rokas *et al.* 2003; Hassanin 2006; Philippe *et al*. 2005).

Use appropriate alignment algorithms and check alignments manually

As discussed above, large multigene datasets make particular demands on alignment algorithms, and these should be taken into account when designing phylogenetic studies. Pending the further development of automatic alignment quality evaluation, alignments should be checked manually before being included in analyses.

Explore taxon selection

The work described in this thesis has shown that taxon selection can be a crucial influence in the phylogenetic conclusions drawn from an analysis (Graybeal 1998; Slack *et al.* 2006; Wiens 2005; Wiens 2006). Researchers should explore the effects of including different taxa and varying numbers of representatives of higher groups in alignments. In particular, researchers should investigate the effect of increasing taxon sampling, despite the likely presence of missing data in large datasets (Wiens 2003).

Explore model choice

In both analyses described in this thesis, model choice has affected phylogenetic conclusions. Simulation studies back up this conclusion (Lemmon and Moriarty 2004). Models should be tested in a statistical Bayesian framework to determine their goodness-of-fit (Nylander *et al.* 2004). Carrying out phylogenetic analysis in a Bayesian framework allows non-nested, complex models to be compared using Bayes Factors (Kass and Raftery 1995) and exploration of the effects of different models

should be a part of any large-scale phylogenetic analysis. In particular, partitioning schemes should be used when multiple genes are involved.

Look for systematic bias

With the large number of characters involved, multigene studies are particularly susceptible to systematic biases. Well-known sources of bias such as base composition, long branches, and mitochondrial strand bias should be investigated and, if found, eliminated.

The outlook for multigene phylogenetics is bright. Increasing production of sequence data, and increasing availability of computing power can be taken as a given. To make the most of these resources, automated phylogenetics tools will be necessary, along with developments in alignment, tree reconstruction, phylogenetic models and orthology clustering. Efforts like TaxMan, TreeBase (Sanderson *et al.* 1994) and AMIGA (Feijao *et al.* 2006) represent the first step in this direction. The lessons learned from multigene studies, such as those described in this thesis, will allow the next generation of tools to better fit the needs of modern phylogenetics research.

Bibliography

Abascal F, Zardoya R and Posada D, 2005. Prottest: selection of best-fit models of protein evolution. *Bioinformatics* **21**: pp. 2104-2105.

Adoutte, A, Balavoine, G, Lartillot, N, de Rosa, R, 1999. Animal evolution. The end of the intermediate taxa?. *Trends Genet* **15**: 104-108.

Aguinaldo, AM, Turbeville, JM, Linford, LS, Rivera, MC, Garey, J, Raff, RA,Lake, JA, 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature* **387**: 489-493.

Ahola V, Aittokallio T, Vihinen M and Uusipaikka E, 2006. A statistical score for assessing the quality of multiple sequence alignments. *BMC Bioinformatics* **7**: p. 484.

Alexeyenko A, Tamas I, Liu G and Sonnhammer ELL, 2006. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics* **22**: p. e9-15.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ, 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: pp. 3389-3402.

Anderson, FE, Cordoba, AJ and Thollesson, M, 2004. Bilaterian phylogeny based on analyses of a region of the sodium-potassium ATPase beta-subunit gene. *J Mol Evol* **58**: 252-268.

Balavoine, G, 1997. The early emergence of platyhelminths is contradicted by the agreement between 18S rRNA and Hox genes data. *C R Acad Sci III* **320**: 83-94.

Baldauf, SL and Palmer, JD, 1993. Animals and fungi are each other's closest relatives: congruent evidence from multiple proteins. *Proc Natl Acad Sci U S A* **90**: 11558-11562.

Bateman A, Birney E, Cerruti L, Durbin R, Etwiller L, Eddy SR, Griffiths-Jones S, Howe KL, Marshall M and Sonnhammer ELL, 2002. The PFAM protein families database. *Nucleic Acids Res* **30**: pp. 276-280.

Benson, DA, Karsch-Mizrachi, I, Lipman, DJ, Ostell, J and Wheeler, DL, 2006. GenBank. *Nucleic Acids Res* **34**: D16-20.

Bernardi G, Olofsson B, Filipski J, Zerial M, Salinas J, Cuny G, Meunier-Rotival M and Rodier F, 1985. The mosaic genome of warm-blooded vertebrates. *Science* **228**: pp. 953-958

Berney C, Pawlowski J and Zaninetti L, 2000. Elongation factor 1-alpha sequences do not support an early divergence of the Acoela. *Mol Biol Evol* **17**: pp. 1032-1039.

Bininda-Emonds, ORP, 2004a. Trees versus characters and the supertree/supermatrix "paradox". *Syst Biol* **53**: 356-359.

Bininda-Emonds, ORP, 2004b. The evolution of supertrees. *Trends Ecol Evol* **19**: 315-322.

Black WC 4th and Roehrdanz RL, 1998. Mitochondrial gene order is not conserved in arthropods: prostriate and metastriate tick mitochondrial genomes. *Mol Biol Evol* **15**: pp. 1772-1785.

Blackledge TA and Gillespie RG, 2004. Convergent evolution of behavior in an adaptive radiation of hawaiian web-building spiders. *Proc Natl Acad Sci U S A* **101**: pp. 16228-16233.

Blair JE, Ikeo K, Gojobori T and Hedges SB, 2002. The evolutionary position of nematodes. *BMC Evol Biol* **2**: p. 7.

Boguski MS, Lowe TM and Tolstoshev CM, 1993. DBEST - database for "expressed sequence tags". *Nat Genet* **4**: pp. 332-333.

Brandley MC, Schmitz A and Reeder TW, 2005. Partitioned Bayesian analyses, partition choice, and the phylogenetic relationships of scincid lizards. *Syst Biol* **54**: pp. 373-390.

Brinkmann H, Giezen M, Zhou Y, Raucourt G and Philippe H, 2005. An empirical assessment of long-branch attraction artefacts in deep eukaryotic phylogenomics.. *Syst Biol* **54**: p. 743--757.

Bruno WJ, Socci ND and Halpern AL, 2000. Weighted neighbor joining: a likelihood-based approach to distance-based phylogeny reconstruction. *Mol Biol Evol* **17**: pp. 189-197.

Burge C and Karlin S, 1997. Prediction of complete gene structures in human genomic dna. *J Mol Biol* **268**: pp. 78-94.

Cadez N, Raspor P and Smith MT, 2006. Phylogenetic placement of *hanseniaspora-kloeckera* species using multigene sequence analysis with taxonomic implications: descriptions of *hanseniaspora pseudoguilliermondii sp. nov.* and *hanseniaspora occidentalis var. citrica var. nov. Int J Syst Evol Microbiol* **56**: pp. 1157-1165.

Cameron S, Barker S and Whiting M, 2006. Mitochondrial genomics and the new insect order mantophasmatodea.. *Mol Phylogenet Evol* **38**: p. 274--279.

Carranza, S, Baguna, J and Riutort, M, 1997. Are the Platyhelminthes a monophyletic primitive group? An assessment using 18S rDNA sequences. *Mol Biol Evol* **14**: 485-497.

Castresana J, 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Mol Biol Evol* **17**: pp. 540-552.

Chen, D, Eulenstein, O, Fernandez-Baca, D and Sanderson, M, 2006. Minimum-flip supertrees: complexity and algorithms. *IEEE/ACM Trans Comput Biol Bioinform* **3**: 165-173.

Chiu JC, Lee EK, Egan MG, Sarkar IN, Coruzzi GM and DeSalle R, 2006. OrthologID: automation of genome-scale ortholog identification within a parsimony framework. *Bioinformatics* **22**: pp. 699-707.

Clayton DA, 1982. Replication of animal mitochondrial DNA. *Cell* **28**: pp. 693-705.

Cole J, Chai B, Farris R, Wang Q, Kulam-Syed-Mohideen A, McGarrell D, Bandela A, Cardenas E, Garrity G and Tiedje J, 2006. The ribosomal database project (RDP-II): introducing myRDP space and quality controlled public data. *Nucleic Acids Res* .doi: 10.1093/nar/gkl889

Cook CE, Yue Q and Akam M, 2005. Mitochondrial genomes suggest that hexapods and crustaceans are mutually paraphyletic. *Proc Biol Sci* **272**: pp. 1295-1304.

Crease TJ, 1999. The complete sequence of the mitochondrial genome of *Daphnia pulex* (Cladocera: Crustacea). *Gene* **233**: pp. 89-99.

Damen WGM, Janssen R and Prpic N, 2005. Pair rule gene orthologs in spider segmentation. *Evol Dev* **7**: pp. 618-628.

Dávila S, Piñero D, Bustos P, Cevallos M and Dávila G, 2005. The mitochondrial genome sequence of the scorpion *centruroides limpidus* (karsch 1879) (Chelicerata; Arachnida).. *Gene* **360**: p. 92--102.

de la Torre JEB, Egan MG, Katari MS, Brenner ED, Stevenson DW, Coruzzi GM and DeSalle R, 2006. Estimating plant phylogeny: lessons from partitioning. *BMC Evol Biol* **6**: p. 48.

de Queiroz A and Gatesy, J, 2006. The supermatrix approach to systematics. *Trends Ecol Evol* : doi:10.1016/j.tree.2006.10.002

Dopazo, H and Dopazo, J, 2005. Genome-scale evidence of the nematode-arthropod clade. *Genome Biol* **6**: R41.

Dotson EM and Beard CB, 2001. Sequence and organization of the mitochondrial genome of the chagas disease vector, *Triatoma dimidiata*. *Insect Mol Biol* **10**: pp. 205-215.

Driskell AC, Ane C, Burleigh JG, McMahon MM, O'meara BC and Sanderson MJ, 2004. Prospects for building the tree of life from large sequence databases. *Science* **306**: pp. 1172-1174.

Drummond A and Rambaut A, 2003. BEAST. Available from http://evolve.zoo.ox.ac.uk/beast/

Dufayard J, Duret L, Penel S, Gouy M, Rechenmann F and Perriere G, 2005. Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics* **21**: pp. 2596-2603.

Dunlop J and Braddy S, 2001. Scorpions and their sister-group relationships. In *Scorpions 2001. In Memoriam Gary A Polis*. Fet V & Selden PA (Eds.). Burnham beeches, UK: The British Arachnological Society pp. 1-24.

Eddy S, 2001. Hmmer: profile hidden markov models for biological sequence analysis. Washington University School of Medicine, St Louis, MO

Edgar RC, 2004. Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: pp. 1792-1797.

Enright AJ, Van Dongen S and Ouzounis CA, 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30**: pp. 1575-1584.

Erixon P, Svennblad B, Britton T and Oxelman B, 2003. Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. *Syst Biol* **52**: 665-673.

Eulenstein O, Chen D, Burleigh JG, Fernandez-Baca D and Sanderson MJ, 2004. Performance of flip supertree construction with a heuristic algorithm. *Syst Biol* **53**: 299-308.

Farris JS, Albert VA, Källersjö M, Lipscomb D and Kluge AG, 1996. Parsimony Jackknifing outperforms Neighbor-Joining. *Cladistics* **12**: 99–124.

Feijao PC, Neiva LS, de Azeredo-Espin AML and Lessinger AC, 2006. AMiGA: the arthropodan mitochondrial genomes accessible database. *Bioinformatics* **22**: pp. 902-903.

Felsenstein J, 1978. Cases in which parsimony or compatibility methods will be positively misleading. *Syst Zool* **27**: pp. 401-410.

Felsenstein J, 1985. Confidence-limits on phylogenies - an approach using the bootstrap. *Evolution* **39**: pp. 783-791.

Felsenstein J, 2005. Phylip (phylogeny inference package) version 3.6. Distributed by the author. Department of Genome Sciences, University of Washington, Seattle.

Foster PG and Hickey DA, 1999. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. J Mol Evol **48**: pp. 284-290.

Fukunishi Y and Hayashizaki Y, 2001. Amino acid translation program for full-length cdna sequences with frameshift errors. *Physiol Genomics* **5**: pp. 81-87.

Funch P and Kristensen RM, 1995. Cycliophora is a new phylum with affinities to Entoprocta and Ectoprocta. *Nature* **378**: p. 711–14.

Fundel K and Zimmer R, 2006. Gene and protein nomenclature in public databases. *BMC Bioinformatics* **7**: p. 372.

Gadagkar S and Kumar S, 2005. Maximum likelihood outperforms maximum parsimony even when evolutionary rates are heterotachous.. *Mol Biol Evol* **22**: p. 2139--2141.

Galtier N and Gouy M, 1995. Inferring phylogenies from DNA sequences of unequal base compositions. *Proc Natl Acad Sci U S A* **92**: p. 11317--11321.

Gantenbein B and Largiadèr C, 2003. The phylogeographic importance of the strait of Gibraltar as a gene flow barrier in terrestrial arthropods: a case study with the scorpion *Buthus occitanus* as model organism.. *Mol Phylogenet Evol* **28**: p. 119--130.

Garey JR, Near TJ, Nonnemacher MR and Nadler SA, 1996. Molecular evidence for Acanthocephala as a subtaxon of Rotifera. *J Mol Evol* **43**: pp. 287-292.

Gatesy J, Baker RH and Hayashi C, 2004. Inconsistencies in arguments for the supertree approach: supermatrices versus supertrees of Crocodylia. *Syst Biol* **53**: 342-355.

Giribet G, 2003. Molecules, development and fossils in the study of metazoan evolution; Articulata versus Ecdysozoa revisited. *Zoology (Jena)* **106**: pp. 303-326.

Giribet G, Distel DL, Polz M, Sterrer W and Wheeler WC, 2000. Triploblastic relationships with emphasis on the acoelomates and the position of Gnathostomulida, Cycliophora, Plathelminthes, and Chaetognatha: a combined approach of 18s rDNA sequences and morphology. *Syst Biol* **49**: pp. 539-562.

Giribet G, Edgecombe G, Wheeler W and Babbitt C, 2002. Phylogeny and systematic position of Opiliones: a combined analysis of chelicerate relationships using morphological and molecular data.. *Cladistics* **18**: p. 5--70.

Giribet G, Richter S, Edgecombe G and Wheeler W, 2005. The position of crustaceans within Arthropoda — evidence from nine molecular loci and morphology. *Crustacean Issues* **16**: p. 307–352.

Goodstadt L and Ponting C, 2006. Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. *PLoS Comput Biol* **2**: e133

Gordon D, Abajian C and Green P, 1998. Consed: a graphical tool for sequence finishing. *Genome Res* **8**: pp. 195-202.

Graybeal A, 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? *Syst Biol* **47**: pp. 9-17.

Guindon S and Gascuel O, 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Syst Biol* **52**: pp. 696-704.

Halanych K, 2004. The new view of animal phylogeny. *Ann Rev Eco Evol Syst* **35**: pp. 229-256.

Halanych KM, Bacheller JD, Aguinaldo AM, Liva SM, Hillis DM, Lake JA, 1995. Evidence from 18S ribosomal DNA that the lophophorates are protostome animals. *Science* **267**: 1641-1643.

Harris JD, 2003. Can you bank on GenBank?. *Trends Eco Evol* **18**: pp. 317-319.

Hassanin A, 2006. Phylogeny of Arthropoda inferred from mitochondrial sequences: strategies for limiting the misleading effects of multiple changes in pattern and rates of substitution. *Mol Phylogenet Evol* **38**: pp. 100-116.

Hassanin A, Leger N and Deutsch J, 2005. Evidence for multiple reversals of asymmetric mutational constraints during the evolution of the mitochondrial genome of metazoa, and consequences for phylogenetic inferences. *Syst Biol* **54**: pp. 277-298.

Herlyn H, Piskurek O, Schmitz J, Ehlers U and Zischler H, 2003. The syndermatan phylogeny and the evolution of acanthocephalan endoparasitism as inferred from 18s rDNA sequences. *Mol Phylogenet Evol* **26**: pp. 155-164.

Higgins DG, Bleasby AJ and Fuchs R, 1992. CLUSTALV: improved software for multiple sequence alignment. *Comput Appl Biosci* **8**: pp. 189-191.

Huelsenbeck J and Rannala B, 2004. Frequentist properties of Bayesian posterior probabilities of phylogenetic trees under simple and complex substitution models. *Syst Biol* **53**: 904-913.

Hughes J, Longhorn SJ, Papadopoulou A, Theodorides K, de Riva A, Mejia-Chang M, Foster PG and Vogler AP, 2006. Dense taxonomic EST sampling and its applications for molecular systematics of the Coleoptera (beetles). *Mol Biol Evol* **23**: pp. 268-278.

Iseli C, Jongeneel CV and Bucher P, 1999. Estscan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. *Proc Int Conf Intell Syst Mol Biol* : pp. 138-148.

Janies DA and Wheeler WC, 2002. Theory and practice of parallel direct optimization. *EXS* : pp. 115-123.

Jones M and Blaxter M, 2005. Evolutionary biology: animal roots and shoots. *Nature* **434**: pp. 1076-1077.

Kass R and Raftery A, 1995. Bayes factors. *Journal of the American Statistical Association* **90**: pp. 773-795.

Kearney M, 2002. Fragmentary taxa, missing data, and ambiguity: mistaken assumptions and conclusions. *Syst Biol* **51**: pp. 369-381.

Knight RD, Freeland SJ and Landweber LF, 2001. A simple model based on mutation and selection explains trends in codon and amino-acid usage and GC composition within and across genomes. Genome Biol **2**: RESEARCH0010.

Kolaczkowski B and Thornton JW, 2004. Performance of maximum parsimony and likelihood phylogenetics when evolution is heterogeneous. *Nature* **431**: pp. 980-984.

Koonin EV, 2005. Orthologs, paralogs, and evolutionary genomics. *Annu Rev Genet* **39**: pp. 309-338.

Kouprina N, Pavlicek A, Mochida GH, Solomon G, Gersch W, Yoon Y, Collura R, Ruvolo M, Barrett JC, Woods CG, Walsh CA, Jurka J and Larionov V, 2004. Accelerated evolution of the aspm gene controlling brain size begins prior to human brain expansion. *PLoS Bio*l **2**: p. E126

Lavrov D, Boore J and Brown W, 2000. The complete mitochondrial dna sequence of the horseshoe crab *Limulus polyphemus*. *Mol Biol Evol* **17**: p. 813—824.

Lee C, Grasso C and Sharlow MF, 2002. Multiple sequence alignment using partial order graphs. *Bioinformatics* **18**: pp. 452-464.

Lemmon AR and Moriarty EC, 2004. The importance of proper model assumption in Bayesian phylogenetics. *Syst Biol* **53**: pp. 265-277.

Li L, Stoeckert CJJ and Roos DS, 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res* **13**: pp. 2178-2189.

Littlewood D, Rohde K and Clough KA, 1999. The interrelationships of all major groups of Platyhelminthes: phylogenetic evidence from morphology and molecules. *Bio J Lin Soc* **66**: pp. 75-114.

Lockhart PK, Steel MA, Hendy MD,Penny D, 1994. Recovering Evolutionary Trees under a More Realistic Model of Sequence. *Mol Biol Evol* **11**: 605-612.

Ludwig W, Strunk O, Westram R, Richter L, Meier H, Buchner A, Lai T, Steppi S, Jobb G, Forster W, Brettske I, Gerber S, Ginhart AW, Gross O, Grumann S, Hermann S, Jost R, Konig A, Liss T, Lussmann R, May M, Nonhoff B, Reichel B, Strehlow R, Stamatakis A, Stuckmann N, Vilbig A, Lenke M, Ludwig T, Bode A and Schleifer K, 2004. ARB: a software environment for sequence data. *Nucleic Acids Res* **32**: pp. 1363-1371.

Maddison DR, Swofford DL and Maddison WP, 1997. Nexus: an extensible file format for systematic information. *Syst Biol* **46**: pp. 590-621.

Mallatt J and Giribet G, 2006. Further use of nearly complete 28s and 18s rRNA genes to classify Ecdysozoa: 37 more arthropods and a kinorhynch. *Mol Phylogenet Evol* **40**: pp. 772-794.

Mallatt J and Winchell CJ, 2002. Testing the new animal phylogeny: first use of combined large-subunit and small-subunit rRNA gene sequences to classify the protostomes. *Mol Biol Evol* **19**: pp. 289-301.

Mallatt JM, Garey JR and Shultz JW, 2004. Ecdysozoan phylogeny and Bayesian inference: first use of nearly complete 28s and 18s rRNA gene sequences to classify the arthropods and their kin. *Mol Phylogenet Evol* **31**: pp. 178-191.

Manuel M, Jager M, Murienne J, Clabaut C and Guyader HL, 2006. Hox genes in sea spiders (Pycnogonida) and the homology of arthropod head segments. *Dev Genes Evol* **216**: pp. 481-491.

Martin DMA, Berriman M and Barton GJ, 2004. Gotcha: a new method for prediction of protein function assessed by the annotation of seven genomes. *BMC Bioinf* **5**: p. 178.

Maxmen A, Browne W, Martindale M and Giribet G, 2005. Neuroanatomy of sea spiders implies an appendicular origin of the protocerebral segment.. *Nature* **437**: p. 1144--1148.

Medina M, Collins AG, Silberman JD and Sogin ML, 2001. Evaluating hypotheses of basal animal phylogeny using complete sequences of large and small subunit rRNA. *Proc Natl Acad Sci U S A* **98**: pp. 9707-9712.

Morris SC, Cohen BL, Gawthrop AP, Cavalier-Smith T, Winnepenninckx B, 1996. Lophophorate phylogeny. *Science* **272**: 282-283.

Nardi F, Spinsanti G, Boore J, Carapelli A, Dallai R and Frati F, 2003. Hexapod origins: monophyletic or paraphyletic?. *Science* **299**: p. 1887--1889.

Navajas M, Le Conte Y, Solignac M, Cros-Arteil S and Cornuet J, 2002. The complete sequence of the mitochondrial genome of the honeybee ectoparasite mite *Varroa destructor* (Acari: Mesostigmata). *Mol Biol Evol* **19**: pp. 2313-2317.

Negrisolo E, Minelli A and Valle G, 2004. Extensive gene order rearrangement in the mitochondrial genome of the centipede *Scutigera coleoptrata*. *J Mol Evol* **58**: pp. 413-423.

Nielsen C, 1995. *Interrelationships of the living phyla*. Oxford University Press, Great Clarendon Street, Oxford

Nylander JAA, Ronquist F, Huelsenbeck JP and Nieves-Aldrey JL, 2004. Bayesian phylogenetic analysis of combined data. *Syst Biol* **53**: pp. 47-67.

Nylander, J.A.A., 2004. MrModeltest v2. Program distributed by the author. Evolutionary Biology Centre, Uppsala University.

Oliveira GC and Kemp WM, 1995. Cloning of two actin genes from *Schistosoma mansoni*. *Mol Biochem Parasitol* **75**: pp. 119-122.

Olson PD and Tkach VV, 2005. Advances and trends in the molecular systematics of the parasitic Platyhelminthes. *Adv Parasitol* **60**: pp. 165-243.

Olson SA, 2002. Emboss opens up sequence analysis. european molecular biology open software suite. *Brief Bioinform* **3**: pp. 87-91.

Page RDM and Valiente G, 2005. An edit script for taxonomic classifications. *BMC Bioinf* **6**: p. 208.

Page RDM, 2005. A taxonomic search engine: federating taxonomic databases using web services. *BMC Bioinf* **6**: p. 48.

Pagel M and Meade A, 2004. A phylogenetic mixture model for detecting pattern-heterogeneity in gene sequence or character-state data. *Syst Biol* **53**: pp. 571-581.

Parkinson J, Anthony A, Wasmuth J, Schmid R, Hedley A and Blaxter M, 2004a. Partigene--constructing partial genomes. *Bioinformatics* **20**: pp. 1398-1404.

Parkinson J, Whitton C, Schmid R, Thomson M and Blaxter M, 2004b. Nembase: a resource for parasitic nematode ESTs. *Nucleic Acids Res* **32**: p. D427-30.

Passamaneck Y and Halanych KM, 2006. Lophotrochozoan phylogeny assessed with LSU and SSU data: evidence of lophophorate polyphyly. *Mol Phylogenet Evol* **40**: pp. 20-28.

Pertea G, Huang X, Liang F, Antonescu V, Sultana R, Karamycheva S, Lee Y, White J, Cheung F, Parvizi B, Tsai J and Quackenbush J, 2003. Tigr gene indices clustering tools (tgicl): a software system for fast clustering of large EST datasets. *Bioinformatics* **19**: pp. 651-652.

Peterson KJ and Eernisse DJ, 2001. Animal phylogeny and the ancestry of bilaterians: inferences from morphology and 18s rDNA gene sequences. *Evol Dev* **3**: pp. 170-205.

Petrov NB, Vladychenskaia,NS, 2005. Phylogeny of protostome moulting animals (Ecdysozoa) inferred from 18 and 28S rRNA gene sequences. *Mol Biol (Mosk)* **39**: 590-601.

Philippe H, 1993. MUST, a computer package of Management Utilities for Sequences and Trees. *Nucleic Acids Res* **21**: 5264-5272

Philippe H, Lartillot N and Brinkmann H, 2005. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. *Mol Biol Evol* **22**: pp. 1246-1253.

Philippe H, Zhou Y, Brinkmann H, Rodrigue N and Delsuc F, 2005. Heterotachy and long-branch attraction in phylogenetics.. *BMC Evol Biol* **5**: p. 50.

Philippe H, Snell EA, Bapteste E, Lopez P, Holland PWH and Casane D, 2004. Phylogenomics of eukaryotes: impact of missing data on large alignments. *Mol Biol Evol* **21**: pp. 1740-1752.

Posada D and Buckley TR, 2004. Model selection and model averaging in phylogenetics: advantages of Akaike Information Criterion and Bayesian approaches over likelihood ratio tests. *Syst Biol* **53**: 793-808.

Posada D, Crandall K, 1998. MODELTEST: testing the model of DNA substitution.. *Bioinformatics* **14**: 817--818.

Pupko T, Huchon D, Cao Y, Okada N and Hasegawa M, 2002. Combining multiple data sets in a likelihood analysis: which models are the best?. *Mol Biol Evol* **19**: p. 2294--2307.

Qiu Y, Song D, Zhou K and Sun H, 2005. The mitochondrial sequences of *Heptathela hangzhouensis* and *Ornithoctonus huwena* reveal unique gene arrangements and atypical tRNAs. *J Mol Evol* **60**: pp. 57-71.

Regier JC and Shultz JW, 2001. Elongation factor-2: a useful gene for arthropod phylogenetics. *Mol Phylogenet Evol* **20**: pp. 136-148.

Regier JC, Shultz JW and Kambic RE, 2005. Pancrustacean phylogeny: hexapods are terrestrial crustaceans and maxillopods are not monophyletic. *Proc Biol Sci* **272**: pp. 395-401.

Rokas A and Carroll SB, 2005. More genes or more taxa? the relative contribution of gene number and taxon number to phylogenetic accuracy. *Mol Biol Evol* **22**: pp. 1337-1344.

Rokas A, Williams BL, King N and Carroll SB, 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* **425**: pp. 798-804.

Ronquist F and Huelsenbeck J, 2003. Mrbayes 3: Bayesian phylogenetic inference under mixed models.. *Bioinformatics* **19**: p. 1572--1574.

Ruiz-Trillo I, Riutort M, Fourcade HM, Baguna J and Boore JL, 2004. Mitochondrial genome data support the basal position of Acoelomorpha and the polyphyly of the Platyhelminthes. *Mol Phylogenet Evol* **33**: pp. 321-332.

Sanderson M, Donoghue M, Piel W and Eriksson T, 1994. Treebase: a prototype database of phylogenetic analyses and an interactive tool for browsing the phylogeny of life. http://www.treebase.org/treebase/

Sanderson MJ and Driskell AC, 2003. The challenge of constructing large phylogenetic trees. *Trends Plant Sci* **8**: pp. 374-379.

Sanderson MJ, Driskell AC, Ree RH, Eulenstein O and Langley S, 2003. Obtaining maximal concatenated phylogenetic data sets from large sequence databases. *Mol Biol Evol* **20**: pp. 1036-1042.

Schmidt HA, Strimmer K, Vingron M and von Haeseler A, 2002. Tree-puzzle: maximum likelihood phylogenetic analysis using quartets and parallel computing. *Bioinformatics* **18**: pp. 502-504.

Schoppmeier M and Damen WGM, 2005. Suppressor of hairless and presenilin phenotypes imply involvement of canonical notch-signalling in segmentation of the spider *Cupiennius salei. Dev Biol* **280**: pp. 211-224.

Schultz J, 1990. Evolutionary morphology and phylogeny of Arachnida. *Cladistics* **6**: p. 1-38.

Shao R, Mitani H, Barker SC, Takahashi M and Fukunaga M, 2005. Novel mitochondrial gene content and gene arrangement indicate illegitimate inter-mtDNA recombination in the chigger mite, *Leptotrombidium pallidum. J Mol Evol* **60**: pp. 764-773.

Siveter DJ, Sutton MD, Briggs DEG and Siveter DJ, 2004. A silurian sea spider. *Nature* **431**: pp. 978-980.

Slack K, Delsuc F, McLenachan P, Arnason U and Penny D, 2007. Resolving the root of the avian mitogenomic tree by breaking up long branches. *Mol Phylogenet Evol* **42** :1-13

Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JGR, Korf I, Lapp H, Lehvaslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD and Birney E, 2002. The BioPerl toolkit: perl modules for the life sciences. *Genome Res* **12**: pp. 1611-1618.

Steenkamp ET, Wright J and Baldauf SL, 2006. The protistan origins of animals and fungi. *Mol Biol Evol* **23**: 93-106.

Sullivan J and Swofford DL, 2001. Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated?. *Syst Biol* **50**: pp. 723-729.

Sutton MD, Briggs DEG, Siveter DJ, Siveter DJ and Orr PJ, 2002. The arthropod *Offacolus kingi* (chelicerata) from the Silurian of Herefordshire, England: computer based morphological reconstructions and phylogenetic affinities. *Proc Biol Sci* **269**: pp. 1195-1203.

Svennblad B, Erixon P, Oxelman B and Britton T, 2006. Fundamental differences between the methods of maximum likelihood and maximum posterior probability in phylogenetics. *Syst Biol* **55**: pp. 116-121.

Swofford DL, 2006. Paup*. phylogenetic analysis using parsimony (*and other methods). version 4.. Sinauer Associates, Sunderland, Massachusetts.

Tanaka M and Ozawa T, 1994. Strand asymmetry in human mitochondrial dna mutations. *Genomics* **22**: p. 327--335.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ and Natale DA, 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics* **4**: p. 41.

Thompson JD, Higgins DG and Gibson TJ, 1994. CLUSTALW: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**: pp. 4673-4680.

Thompson JD, Plewniak F, Ripp R, Thierry JC and Poch O, 2001. Towards a reliable objective function for multiple sequence alignments. *J Mol Biol* **314**: pp. 937-951.

Umetsu K, Iwabuchi N, Yuasa I, Saitou N, Clark PF, Boxshall G, Osawa M and Igarashi K, 2002. Complete mitochondrial DNA sequence of a tadpole shrimp (*Triops Cancriformis*) and analysis of museum samples. *Electrophoresis* **23**: pp. 4080-4084.

Wallace IM, Blackshields G and Higgins DG, 2005. Multiple sequence alignments. *Curr Opin Struct Biol* **15**: pp. 261-266.

Wasmuth JD and Blaxter ML, 2004. Prot4EST: translating expressed sequence tags from neglected genomes. *BMC Bioinformatics* **5**: p. 187.

Wheeler W, Giribet G and Edgecombe G, 2004. Arthropod systematics. In *Assembling The Tree Of Life*. Creacreaft K & Donoghue M (Eds.). Oxford University Press, New York, NY pp. 281-295.

Wheeler WC and Hayashi CY, 1998. The phylogeny of the extant chelicerate orders. *Cladistics* **14**: p. 173–192.

Wiens JJ, 2003. Missing data, incomplete taxa, and phylogenetic accuracy. *Syst Biol* **52**: pp. 528-538.

Wiens JJ, 2005. Can incomplete taxa rescue phylogenetic analyses from long-branch attraction?. *Syst Biol* **54**: pp. 731-742.

Wiens JJ, 2006. Missing data and the design of phylogenetic analyses. *J Biomed Inform* **39**: pp. 34-42.

Wilkinson M, Cotton JA, Creevey C, Eulenstein O, Harris SR, Lapointe F, Levasseur C, McInerney JO, Pisani D. and Thorley JL, 2005. The shape of supertrees to come: tree shape related properties of fourteen supertree methods. *Syst Biol* **54**: 419-431.

Winnepenninckx B, Backeljau T, Mackey LY, Brooks JM, De Wachter R, Kumar S,Garey JR, 1995. 18S rRNA data indicate that Aschelminthes are polyphyletic in origin and consist of at least three distinct clades. *Mol Biol Evol* **12**: 1132-1137.

Wolf YI, Rogozin IB and Koonin EV, 2004. Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis. *Genome Res* **14**: pp. 29-36.

Yan C, Burleigh JG and Eulenstein O, 2005. Identifying optimal incomplete phylogenetic data sets from sequence databases. *Mol Phylogenet Evol* **35**: pp. 528-535.

web reference 1 - The NCBI Entrez database browser.

url: ***http://www.ncbi.nlm.nih.gov/Entrez/***

web reference 2 - The phylota project.

url: ***http://ginger.ucdavis.edu/phylota/***

web reference 3 - The hal project.

url: ***http://aftolorg/pages/Halweb3.htm***

web reference 4 - The treeblaster project.

url: ***http://www.bch.umontreal.ca/pepdb/pepdb_tool.html#TreeBlaster***

web reference 5 - The protist est program (PEP).

url: ***http://www.bch.umontreal.ca/pepdb/pep_main.html***

web reference 6 – The NCBI taxonomy server.

url: ***http://www.ncbi.nlm.nih.gov/Taxonomy/taxonomyhome.html/***

web reference 7 - NCBI taxonomy database dump.

url: ***ftp://ftp.ncbi.nih.gov/pub/taxonomy/taxdump.tar.gz***

web reference 8 - Description of the fasta file format.

url: ***http://www.ncbi.nlm.nih.gov/blast/fasta.shtml***

web reference 9 - Description of the Newick tree file format.

url: ***http://evolution.genetics.washington.edu/phylip/newick_doc.html***

web reference 10 - The perl programming language.

url: ***http://www.perl.com***

web reference 11 - The postgresql relational database management software.

url: ***http://www.postgresql.org/***

web reference 12 - Tracer: A program for analysing results from Bayesian MCMC programs.

url: ***http://evolve.zoo.ox.ac.uk/software.html?id=tracer***

Appendices

1. TaxMan user guide

2. Jones, M and Blaxter, M. TaxMan: a Taxonomic database Manager. Accepted BMC Bioinformatics, December 2006.

3. Jones, M; Gantenbein, B; Fet, V and Blaxter, M. The effect of model choice on phylogenetic inference using mitochondrial sequence data: lessons from the scorpions. In press Molecular Phylogenetics and Evolution, November 2006.